IFW

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/063,557	05/02/2002	Audrey Goddard	GNE.3230R1C39	9770

7590 06/09/2005

AnneMarie Kaiser
Knobbe Martens Olson & Bear
Sixteenth Floor
620 Newport Center Drive
Newport Beach, CA 92660

EXAMINER	
BLANCHARD, DAVID J	
ART UNIT	**PAPER NUMBER**
1642	

DATE MAILED: 06/09/2005

Please find below and/or attached an Office communication concerning this application or proceeding.

RECEIVED
OIPE/IAP
JUN 21 2005

Office Action Summary	Application No.	Applicant(s)	
	10/063,557	EATON ET AL.	
	Examiner	Art Unit	
	David J. Blanchard	1642	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If the period for reply specified above is less than thirty (30) days, a reply within the statutory minimum of thirty (30) days will be considered timely.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *22 March 2005*.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-5* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-5* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☒ Information Disclosure Statement(s) (PTO-1449 or PTO/SB/08) Paper No(s)/Mail Date *3/22/05; 4/13/05*.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____.
5) ☐ Notice of Informal Patent Application (PTO-152)
6) ☒ Other: *Exhibit A*.

MC

DETAILED ACTION

1. A request for continued examination under 37 CFR 1.114, including the fee set forth in 37 CFR 1.17(e), was filed in this application after final rejection. Since this application is eligible for continued examination under 37 CFR 1.114, and the fee set forth in 37 CFR 1.17(e) has been timely paid, the finality of the previous Office action has been withdrawn pursuant to 37 CFR 1.114. Applicant's submission filed on 22 March 2005 has been entered.

2. Claim 6 has been canceled.

Claim 1 has been amended.

3. Claims 1-5 are pending and under examination.

4. The text of those sections of Title 35, U.S. Code not included in this action can be found in a prior Office action.

5. This Office Action contains New Grounds of Rejections.

Inventorship

6. The request for the deletion of inventors Eaton, Filvaroff, Gerristen and Watanabe is approved and the inventors have been deleted.

Rejections Withdrawn

7. The rejection of claims 1-5 under 35 U.S.C. 103(a) as being unpatentable over Lal et al (WO 00/00610, 1/6/2000, cited previously) in view of Queen et al (US Patent

5,530,101, issued 6/1996) is withdrawn in view of applicants arguments and the fact that Lal et al do not teach the expression of the polypeptide or a function for the protein.

Response to Arguments

8. The rejection of claims 1-5 under 35 U.S.C 101 because the claimed invention is not supported by a substantial asserted utility or a well-established utility is maintained.

The response filed 3/22/2005 has been carefully considered, but is deemed not to be persuasive. Applicant reviews the evidentiary standard regarding the legal presumption of utility. The examiner takes no issue with Applicant's discussion of the evidentiary standard regarding the legal presumption of utility. Applicant argues that the utility need not be proved to a statistical certainty, a reasonable correlation between the evidence and the asserted utility is sufficient and applicant cites numerous case law in support of applicants arguments that for a therapeutic and diagnostic use, utility does not have to be established to an absolute certainty and the evidence need not be direct evidence so long as there is a reasonable correlation between the evidence and the asserted utility. Applicant argues that as set forth in MPEP 2107 II(B)(1) "If applicant has asserted that the claimed invention is useful for any particular practical purpose... and the assertion would be considered credible by a person of ordinary skill in the art, do not impose a rejection based on lack of utility." In response to these arguments, the examiner agrees with Applicant's statement that absolute certainty is not the legal standard for utility. However, the rejection does not question the presumption of truth, or credibility, of the asserted utility. The asserted utilities of cancer diagnostics and

cancer therapeutics for the claimed polypeptides are credible and specific, however, they are not substantial. The data set forth in the specification are preliminary at best because the specification does not teach the expression of the PRO1069 polypeptide nor any particular biological activity of the polypeptide. Applicant summarizes their arguments and the disputed issues involved. Applicant reiterates that Example 18 in the specification shows that mRNA encoding the PRO1069 polypeptide is more highly expressed in normal kidney compared to kidney tumor and applicant asserts that it is well-established in the art that a change in the level of mRNA for a particular protein, generally leads to a corresponding change in the level of the encoded protein and based on the identification of the mRNA encoding the PRO1069 polypeptide under-expressed in tumor tissue compared to normal tissue renders the PRO1069 polypeptide useful as a diagnostic tool for the determination of the presence or absence of tumor. In support, applicant again argues with the declaration of J. Christopher Grimaldi (previously submitted as Exhibit 1) that there is at least a two-fold difference in PRO1069 mRNA between kidney tumor and normal kidney tissue. This has been fully considered, but is not found persuasive. First, it is important to note that the instant specification provides no information regarding PRO1069 polypeptide levels in tumor samples relative to normal samples. Only gene expression data was presented. Therefore, the declaration is insufficient to overcome the rejection of claims 4-9, 11-17 based upon 35 U.S.C. 101 and 112, first paragraph, since it is limited to a discussion of data regarding the gene expression of the PRO1069 cDNA and not gene expression levels and polypeptide levels. Furthermore, the declaration does not provide data such

that the examiner can independently draw conclusions. There is no evidentiary support to Dr. Grimaldi's statement that if a difference in gene expression is detected, this indicates that the gene and its corresponding polypeptide and antibodies against the polypeptide are useful for diagnostic purposes, to screen samples to differentiate between normal and tumor. Finally, it is noted that the literature cautions researchers from drawing conclusions based on small changes in transcript expression levels between normal and cancerous tissue. For example, Hu et al (Journal of Proteome Research 2:405-412, 2003, Ids reference 23 filed 3/31/2005) analyzed 2286 genes that showed a greater than 1-fold difference in mean expression level between breast cancer samples and normal samples in a micoarray (p. 408, middle of right column). Hu et al. discovered that, for genes displaying a 5-fold change or less in tumors compared to normal, there was no evidence of a correlation between altered gene expression and a known role in the disease. However, among genes with a 10-fold or more change in expression level, there was a strong and significant correlation between expression level and a published role in the disease (see discussion section).

Applicant argues that they have established that the accepted understanding in the art is that there is a direct correlation between mRNA levels and the level of expression of the encoded protein and applicant argues with the previously submitted second declaration of J. Christopher Grimaldi (previously submitted as Exhibit 2), which states that those who work in this field are well aware that in the vast majority of cases, when a gene is over-expressed ...the gene product or polypeptide will also be over-expressed and this same principle applies to gene under-expression. Further, applicant

argues with the declaration of Dr. Paul Polakis (previously submitted as Exhibit 3) which states that based upon his experience accumulated in more than 20 years of research, that it is his scientific opinion that for human genes, an increased level of mRNA in a tumor cell relative to a normal cell typically correlates to a similar increase of the encoded protein in the tumor cell relative to the normal cell and that based on his experience although reports exist where such a correlation does not exist, such reports are exceptions to a commonly understood general rule that increased mRNA levels are predictive of corresponding increased levels of the encoded protein and applicant cites Alberts [a] (4th ed. 2002; Exhibit 2), Alberts [b] (3rd ed. 1994; Exhibit 1), Lewin and Zhigang for support that mRNA expression correlates with protein expression. The declarations of Dr. Grimaldi and Dr. Polakis and applicant's arguments have been fully considered, but are not found persuasive. Alberts [b] and Lewin actually support the fact that further research would have to be carried out to determine if the polypeptide expression levels track with the expression levels of the corresponding mRNA. Alberts and Lewin show that there are several levels that control gene expression both at the transcriptional (i.e., mRNA synthesis) and the translational (i.e., protein production) levels. Thus, one skilled in the art would not accept that increased mRNA levels directly correlate with the level of the corresponding polypeptide in view of the multitude of controls at the transcriptional and translational levels. With respect to applicant's arguments regarding the art of Zhigang et al, the art of Zhigang et al does show protein expression, however, the experiments were carried out to demonstrate this and as such Zhigang support that one needs to actually determine the expression of the protein to

be sure of expression. Applicant also argues that Alberts [a] (4th ed. 2002; Exhibit 2). figure 6-3 on page 302 illustrates the general principle that there is a correlation between increased gene expression and increased protein expression. In response to this argument, while increased transcript levels can lead to increased polypeptide levels, there are other regulatory factors that also effect the rate of translation as evidenced by Alberts [b] (Exhibit 1) in Figure 9-72. Additionally, Meric et al (Molecular Cancer Therapeutics, 1:971-979, 2002, Ids reference 17, filed 3/22/2005) teaches that in addition to variations in mRNA sequences that increase or decrease translational efficiency, changes in the expression or availability of components of the translational machinery (i.e., over-expression of eIF4E, eIF4G, eIF-2α, eIF-4A1, ect...) as well as activation of translation through aberrantly activated signal transduction pathways also effect the rate of translation in cancerous cells. Figure 6-3 of Exhibit 2 (Alberts, 4th ed. 2002) does not account for these other types of controls that exist in cancerous cells. Applicant argues that Meric et al states at page 791, left column that the fundamental principle of molecular therapeutics is to exploit differences in gene expression between cancer cells and normal cells and most efforts have concentrated on identifying differences in gene expression at the level of mRNA, which can be attributable to either DNA amplification or to differences in transcription and applicant concludes that those of skill in the art would not be focusing on differences in gene expression between cancer cells and normal cells if there were no correlation between gene expression and protein expression. First, the statements by Meric appear to have been taken out of context. Meric indicates most efforts have concentrated on gene expression at the mRNA level

due to the advent of cDNA array technology, which facilitated this type of analysis. Further, Meric et al in agreement with Alberts and Lewin acknowledges that gene expression is quite complicated and is regulated at the level of mRNA stability, mRNA translation and protein stability and Meric goes on to discuss that the components of the translation machinery and signal pathways involved in the activation of translation initiation represent good targets for cancer therapy (see pages 975-976). If it is the accepted understanding in the art that there is a direct correlation between mRNA levels and the level of expression of the encoded polypeptide, there would not be a need to target the translational machinery, unless of course the two are regulated separately.

Further, applicant argues that the statement of Jang et al (cited previously by the examiner) that "further studies are necessary to determine if changes in protein levels track with changes in mRNA levels for metastasis associated genes in murine tumor cells." does not imply that the reason for additional research is needed is because the levels of mRNA and protein were measured and found not to correlate, rather, the statement simply acknowledges that Jang did not attempt to correlate mRNA and protein levels, and thus further research would be required to do so. In response to this argument, the examiner recognizes that the statement by Jang does not mean that mRNA and protein levels were measured and found not to correlate, the point was the acknowledgement by Jang that further research would be required to determine if a correlation between mRNA and protein levels actually exists. Again, if it is established that the accepted understanding in the art that there is a direct correlation between mRNA levels and the level of expression of the encoded protein, Jang et al would not

state that "further studies are necessary to determine if changes in protein levels track with changes in mRNA levels for metastasis associated genes in murine tumor cells."

Applicant acknowledges that the examiners citations of Vallejo et al, Powell et al, and Fu et al as examples of post-transcriptional regulation of protein levels, they are not inconsistent with applicant's position that mRNA levels correlate, more often than not, with protein levels. In response to this argument, and in agreement with the art of Vallejo et al, Powell et al, Fu et al and Jang et al, Gygi et al (Molecular and Cellular Biology, 19(3):1720-1730, March 1999) states "We found that the correlation between mRNA and protein levels was insufficient to predict protein expression levels from quantitative mRNA data. Indeed, for some genes, while the mRNA levels were of the same value the protein levels varied by more than 20-fold. Conversely, invariant steady-state levels of certain proteins were observed with respective mRNA transcript levels that varied by as much as 30-fold." (see abstract). Also, Haynes et al (1998, Electrophoresis 19:1862-1871, Ids reference 10 filed 3/22/2005), who studied more than 80 proteins relatively homogeneous in half-life and expression level, and found no strong correlation between polypeptide and transcript level. For some genes, equivalent mRNA levels translated into protein abundances, which varied more than 50-fold. Haynes et al concluded that the protein levels cannot be accurately predicted from the level of the corresponding mRNA transcript (p. 1863, second paragraph, and Figure 1). In agreement with Gygi and Haynes, Hanish S. [a] (Nature Reviews, Applied Proteomics Collection, pp. 9-14, March 2005) recently stated "There is a need to profile gene expression at the level of the proteome and to correlate changes in gene-

expression profiles with changes in proteomic profiles. The two are not always linked-numerous alterations occur in protein levels that are not reflected at the RNA level." (see page 12). Further, Hanash [a] teaches that tumors are complex biological systems and no single type of molecular approach fully elucidates tumor behavior, necessitating analysis at multiple levels encompassing genomics and proteomics (see abstract). Hanash et al [b] (The Pharmacogenomics Journal, 3(6):308-311, 2003) states "However perfected DNA microarrays and their analytical tools become for disease profiling, they will not eliminate a pressing need for other types of profiling technologies that go beyond measuring RNA levels, particularly for disease-related investigations." (see page 311). According to Hanash et al [b], there is a need to assay protein levels and activities and numerous alterations may occur in proteins that are not reflected in changes at the RNA level (see page 311). Clearly, contrary to applicant's arguments and as evidenced by the art above, it is not established in the art that the accepted understanding is that there is a direct correlation between mRNA levels and the level of expression of the encoded protein. The literature supports that RNA expression cannot inevitably be correlated with levels of the encoded polypeptide and one skilled in the art would not assume that the levels of RNA are predictive of the levels of the encoded polypeptide given the distinct regulation of transcription and translation as evidenced by Alberts, Lewin, Meric, Jang et al, Vallejo et al, Powell et al, Fu et al, Gygi et al, Haynes et al, Hanash S [a] and Hanash et al [b]. One skilled in the art would do further research to determine whether or not the PRO1069 polypeptide was under-expressed in kidney tumor samples. Such further research requirements make it clear that the

asserted utility is not yet in currently available form, i.e., it is not substantial. This further experimentation is part of the act of invention and until it has been undertaken, Applicant's claimed invention is incomplete. This situation is directly analogous to that which was addressed in *Brenner v. Manson*, 148 U.S.P.Q. 689 (Sup. Ct, 1966), in which the court held that

> "The basic quid pro quo contemplated by the Constitution and the Congress for granting a patent monopoly is the benefit derived by the public from an invention with substantial utility", "[u]nless and until a process is refined and developed to this point-where specific benefit exists in currently available form-there is insufficient justification for permitting an applicant to engross what may prove to be a broad field" and "a patent is not a hunting license" "[i]t is not a reward for the search, but compensation for its successful conclusion."

Applicant refers to three additional articles previously submitted by Applicant (Orntoft et al; Exhibit filed 8/16/2004, Hyman et al; Exhibit filed 8/16/2004, and Pollack et al: Exhibit filed 8/16/2004) as providing evidence that gene amplification generally correlates with levels of the encoded polypeptide. Applicant characterizes Orntoft et al as teaching mRNA and protein levels for individual genes located within amplified or deleted chromosomal regions and found that of the 40 proteins analyzed only one showed disagreement between transcript alteration and protein alteration (Orntoft, page 42). This has been fully considered, but is not found to be persuasive. Orntoft appear to have looked at increased DNA content over large regions of chromosomes and comparing that to mRNA and polypeptide levels from the chromosomal region. This approach to investigating gene copy number was termed CGH. Orntoft et al do not appear to look at gene amplification, mRNA levels and polypeptide levels from a single gene at a time. The instant specification reports data regarding amplification of

individual genes, which may or may not be in a chromosomal region, which is highly amplified. Orntoft et al concentrated on regions of chromosomes with strong gains of chromosomal material containing clusters of genes (page 40). This analysis was not done for PRO1069 in the instant specification. That is, it is not clear whether or not PRO1069 is in a gene cluster in a region of a chromosome that is highly amplified. Therefore, the relevance of Orntoft et al is not clear. Hyman et al used the same CGH approach in their research. Less than half (44%) of highly amplified genes showed mRNA over-expression (abstract). Polypeptide levels were not investigated. Therefore, Hyman et al also do not support utility of the claimed polypeptides. Pollack et al also used CGH technology, concentrating on large chromosome regions showing high amplification (page 12965). Pollack et al did not investigate polypeptide levels. Therefore, Pollack et al also do not support the asserted utility of the claimed invention. Importantly none of the three papers reported that the research was relevant to identifying probes that can be used as cancer diagnostics. The three papers state that the research was relevant to the development of potential cancer therapeutics, but also clearly imply that much further research was needed before such therapeutics were in readily available form. Accordingly, the specifications assertions that the claimed PRO1069 polypeptides have utility in the fields of cancer diagnostics and cancer therapeutics are not substantial.

For these reasons the rejection is maintained.

9. The rejection of claims 1-5 under 35 U.S.C. 112, first paragraph, is maintained. Specifically, since the claimed invention is not supported by a substantial utility or a well-established utility for the reasons set forth above, one skilled in the art clearly would not know how to use the claimed invention.

10. The rejection of claims 1-5 under 35 U.S.C. 112, first paragraph, because the claims contain subject matter, which was not described in the specification in such a way as to enable one skilled in the art to which it pertains, or with which it is most nearly connected, to make and/or use the invention is maintained.

The response filed 3/22/2005 has been carefully considered, but is deemed not to be persuasive. The response argues that in general differential expression levels of mRNA leads to differential protein expression levels and this is the general understanding in the art and the references cited by the examiner are exceptions to the general rule. Applicant relies on example 18 of the specification, the art of Zhigang and Meric for support and states that the totality of the evidence clearly establishes that those of skill in the art would believe that mRNA levels more likely than not correlate with protein levels. In response to this argument and as discussed above in the utility rejection the art of Alberts, Lewin, Meric, Jang et al, Vallejo et al, Powell et al, Fu et al, Gygi et al, Haynes et al, Hanash S [a] and Hanash et al [b] underscores the unpredictability in the art and the predictability of protein translation and its possible use as a diagnostic are not necessarily contingent on the levels of mRNA expression due to the multitude of homeostatic factors affecting transcription and translation. In view of

the totality of evidence of record, one of skill in the art could not predictably use the antibodies of the present claims as a diagnostic or therapeutic agent with a reasonable expectation of success.

New Grounds of Rejections

Priority

Applicant claims priority to five previous applications in the preliminary amendment of 09 September 2002. Priority is granted to PCT/US00/23328, filed 24 August 2000, as the disclosure of '328 is identical to the instant disclosure. However, priority is not granted to USSN 09/380,137, PCT/US99/12252 and 60/088,740 since these applications do not disclose the microarray assay upon which applicant relies for utility of the instantly claimed polypeptides. Therefore, the filing date for the purpose of art rejections is deemed to be 24 August 2000. Applicant is reminded that benefit to a prior-filed application requires written description and enablement under the first paragraph of 35 U.S.C. 112.

Claim Rejections - 35 USC § 102

11. The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

> A person shall be entitled to a patent unless –
>
> (a) the invention was known or used by others in this country, or patented or described in a printed publication in this or a foreign country, before the invention thereof by the applicant for a patent.
>
> (e) the invention was described in (1) an application for patent, published under section 122(b), by another filed in the United States before the invention by the applicant for patent or (2) a patent granted on an application for patent by another filed in the United States before the invention by the applicant for patent,

except that an international application filed under the treaty defined in section 351(a) shall have the effects for purposes of this subsection of an application filed in the United States only if the international application designated the United States and was published under Article 21(2) of such treaty in the English language.

12. Claims 1-2 and 4-5 are rejected under 35 U.S.C. 102(a) as being anticipated by Lal et al (WO 00/00610, 1/6/2000, cited previously on PTO-892 mailed 4/15/2004).

The claims are drawn to an antibody that specifically binds to the polypeptide of SEQ ID NO:50, wherein the antibody is a monoclonal antibody, an antibody fragment and is labeled.

Lal et al teach a polypeptide (SEQ ID NO:35), which is identical to the instantly claimed polypeptide of SEQ ID NO:50 and antibodies that bind the polypeptide are monoclonal, antibody fragments and labeled (see pages 44-45 and 52-53).

13. Claims 1-2 and 4-5 are rejected under 35 U.S.C. 102(e) as being anticipated by Walker et al (U.S. Patent 6,277,574 B1, 4/9/1999).

The claims have been described supra.

Walker et al teach a polypeptide (SEQ ID NO:11) that is identical to the polypeptide of SEQ ID NO:50 (see the alignment attached to the back of this Office Action; Exhibit A) and Walker teaches monoclonal antibodies and antibody fragments that specifically bind the polypeptide and the antibodies may be labeled with a therapeutic agent for treating disease in a subject (see column 13).

14. Claims 1-5 are rejected under 35 U.S.C. 103(a) as being unpatentable over Walker et al (U.S. Patent 6,277,574 B1, 4/9/1999) in view of Queen et al (U.S. Patent 5,530,101, issued 6/96, cited previously on PTO-892 mailed 4/15/2004).

The claims have been described supra. Claim 3 recites wherein the antibody is a humanized antibody.

Walker et al have been described supra. Walker et al does not teach a humanized antibody. This deficiency is made up for in the teachings of Queen et al.

Queen et al teach humanized antibodies for human therapy (see entire document).

It would have been prima facie obvious to one of ordinary skill in the art at the time the claimed invention was made to have produced a humanized antibody to the polypeptide of Walker et al in view of Queen et al.

One of ordinary skill in the art would have been motivated to and had a reasonable expectation of success to have produced a humanized antibody to the polypeptide of Walker et al in view of Queen et al because Walker et al teach the polypeptide of SEQ ID NO:50 (i.e., SEQ ID NO:11 of Walker et al) is associated with kidney disease and it would be obvious in view of Queen et al who teaches humanized antibodies to humanize the antibody of Walker et al for human therapy.

Therefore, the invention as a whole was prima facie obvious to one of ordinary skill in the art at the time the invention was made, as evidenced by the references.

Conclusions

15. No claim is allowed.

16. Any inquiry concerning this communication or earlier communications from the examiner should be directed to David J. Blanchard whose telephone number is (571) 272-0827. The examiner can normally be reached at Monday through Friday from 8:00 AM to 6:00 PM, with alternate Fridays off. If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Jeffrey Siew, can be reached at (571) 272-0787. The official fax number for the organization where this application or proceeding is assigned is 571-273-8300. Any inquiry of a general nature, matching or filed papers or relating to the status of this application or proceeding should be directed to the Kim Downing for Art Unit 1642 whose telephone number is 571-272-0521.

Information regarding the status of an application may be obtained from the patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

Respectfully,
David J. Blanchard
571-272-0827

JEFFREY SIEW
SUPERVISORY PATENT EXAMINER
6/7/05

Exhibit A

```
RESULT 1
US-09-289-349-11
; Sequence 11, Application US/09289349
; Patent No. 6277574
; GENERAL INFORMATION:
;  APPLICANT: Walker, Michael, G.
;  APPLICANT: Volkmuth, Wayne
;  APPLICANT: Klinger, Tod, M.
;  APPLICANT: Azimzai, Yalda
;  APPLICANT: Yue, Henry
;  TITLE OF INVENTION: GENES ASSOCIATED WITH DISEASES OF THE KIDNEY
;  FILE REFERENCE: PB-0010 US
;  CURRENT APPLICATION NUMBER: US/09/289,349
;  CURRENT FILING DATE: 1999-04-09
;  NUMBER OF SEQ ID NOS: 12
;  SOFTWARE: PERL Program
; SEQ ID NO 11
;    LENGTH: 89
;    TYPE: PRT
;    ORGANISM: Homo sapiens
;    FEATURE: -
;    OTHER INFORMATION: 1900433CD1
US-09-289-349-11

  Query Match            100.0%;  Score 461;  DB 3;  Length 89;
  Best Local Similarity  100.0%;  Pred. No. 1.1e-51;
  Matches   89;  Conservative    0;  Mismatches    0;  Indels    0;  Gaps    0;

Qy          1 MERVTLALLLLAGLTALEANDPFANKDDPFYYDWKNLQLSGLICGGLLAIAGIAAVLSGK 60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db          1 MERVTLALLLLAGLTALEANDPFANKDDPFYYDWKNLQLSGLICGGLLAIAGIAAVLSGK 60

Qy         61 CKYKSSQKQHSPVPEKAIPLITPGSATTC 89
              |||||||||||||||||||||||||||||
Db         61 CKYKSSQKQHSPVPEKAIPLITPGSATTC 89
```

PTO/SB/08 Equivalent

INFORMATION DISCLOSURE STATEMENT BY APPLICANT

(Multiple sheets used when necessary)

SHEET 1 OF 2

Application No.	10/063,557
Filing Date	May 2, 2002
First Named Inventor	Goddard, et al.
Art Unit	1642
Examiner	David J. Blanchard
Attorney Docket No.	GNE.3230R1C39

OIPE
MAR 22 2005
PATENT & TRADEMARK OFFICE

U.S. PATENT DOCUMENTS

Examiner Initials	Cite No.	Document Number *Number - Kind Code (if known)* Example: 1,234,567 B1	Publication Date MM-DD-YYYY	Name of Patentee or Applicant	Pages, Columns, Lines Where Relevant Passages or Relevant Figures Appear
DB	1	5,407,799	04-18-1995	Studier	
	2	6,498,235 B2	12-24-2002	Sheppard, et al.	
	3	6,645,499 B1	11-11-2003	Lal, et al.	
DB	4	6,730,502 B2	05-04-2004	Van Hijum, et al.	

FOREIGN PATENT DOCUMENTS

Examiner Initials	Cite No.	Foreign Patent Document *Country Code-Number-Kind Code* Example: JP 1234567 A1	Publication Date MM-DD-YYYY	Name of Patentee or Applicant	Pages, Columns, Lines Where Relevant Passages or Relevant Figures Appear	T[1]

NON PATENT LITERATURE DOCUMENTS

Examiner Initials	Cite No.	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T[1]
DB	5	ALBERTS, et al. 2002. *Molecular Biology of the Cell 4th Edition*, pp. 302, 363-364, 379, 435. New York: Garland Publishing.	
	6	The 1991 Boehringer Mannheim Biochemicals Catalog, page 557, 1991.	
	7	BURGESS, et al. 1990. Possible dissociation of the heparin-binding and mitogenic activities of heparin-binding (acidic fibroblast) growth factor-1 from its receptor-binding activities by site-directed mutagenesis of a single lysine residue. *The Journal of Cell Biology*, 111:2129-2138.	
	8	GRIMALDI, et al. 1989. The t(5;14) chromosomal translocation in a case of acute lymphocytic leukemia joins the interleukin-3 gene to the immunoglobulin heavy chain gene. *Blood*, 73(8):2081-2085.	
DB	9	HANNA, et al. Aug. 1999. HER-2/neu breast cancer predictive testing. *Pathology Associates Medical Laboratories.*	

Examiner Signature	Date Considered
David Blanchard	6/6/05

*Examiner: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

T[1] - Place a check mark in this area when an English language Translation is attached.

PTO/SB/08 Equivalent

INFORMATION DISCLOSURE STATEMENT BY APPLICANT		
	Application No.	10/063,557
	Filing Date	May 2, 2002
	First Named Inventor	Goddard, et al.
	Art Unit	1642
(Multiple sheets used when necessary)	Examiner	David J. Blanchard
SHEET 2 OF 2	Attorney Docket No.	GNE.3230R1C39

OIPE MAR 22 2005 PATENT & TRADEMARK OFFICE

NON PATENT LITERATURE DOCUMENTS			
Examiner Initials	Cite No.	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T[1]
DB	10	HAYNES, et al. 1998. Proteome analysis: Biological assay or data archive? *Electrophoresis*, 19:1862-1871.	
	11	HU, et al. 2003. Analysis of genomic and proteomic data using advanced literature mining. *Journal of Proteome Research*, 2:405-412.	
	12	HYMAN, et al. 2002. Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Research*, 62:6240-6245.	
	13	LAZAR, et al. 1988. Transforming growth factor α: Mutation of aspartic acid 47 and leucine 48 results in different biological activities. *Molecular and Cellular Biology*, 8(3):1247-1252.	
	14	LI, et al. 1980. β-Endorphin omission analogs: Dissociation of immunoreactivity from other biological activities. *Proc. Natl. Acad. Sci. USA*, 77(6):3211-3214.	
	15	LIN, et al. 1975. Structure-function relationships in glucagon: Properties of highly purified Des-His^{1}-, Monoiodo-,and [Des-Asn^{28}, Thr^{29}](homoserine lactone^{27})-glucagon. *Biochemistry*, 14(8):1559-1563.	
	16	MEEKER, et al. 1990. Activation of the interleukin-3 gene by chromosome translocation in acute lymphocytic leukemia with eosinophilia. *Blood*, 76(2):285-289.	
	17	MERIC, et al. 2002. Translation initiation in cancer: A novel target for therapy. *Molecular Cancer Therapeutics*, 1:971-979.	
	18	ØRNTOFT, et al. 2002. Genome-wide study of gene copy numbers, transcripts, and protein levels in pairs of non-invasive and invasive human transitional cell carcinomas. *Molecular & Cellular Proteomics*, 1:37-45.	
	19	POLLACK, et al. 2002. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *PNAS*, 99(20):12963-12968.	
	20	SCHWARTZ, et al. 1987. A superactive insulin: [B10-aspartic acid]insulin(human). *Proc. Natl. Acad. Sci. USA*, 84:6408-6411.	
	21	SINGLETON, et al. 1992. Clinical and pathologic significance of the c-*erb*B-2 (*HER-2/neu*) oncogene. *Pathol. Annu.*, 1(27):165-190.	
DB	22	ZHIGANG, et al. 2004. Prostate stem cell antigen (PSCA) expression in human prostate cancer tissues and its potential role in prostate carcinogenesis and progression of prostate cancer. *World Journal of Surgical Oncology*, 2:13.	

1351252_1:dmb
031805

Examiner Signature	David Blanchard	Date Considered	6/6/05

*Examiner: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

T[1] - Place a check mark in this area when an English language Translation is attached.

PTO/SB/08 Equivalent

INFORMATION DISCLOSURE STATEMENT BY APPLICANT	Application No.	10/063,557
	Filing Date	May 2, 2002
	First Named Inventor	Goddard, et al.
	Art Unit	1642
(Multiple sheets used when necessary)	Examiner	David J. Blanchard
SHEET 1 OF 1	Attorney Docket No.	GNE.3230R1C39

U.S. PATENT DOCUMENTS

Examiner Initials	Cite No.	Document Number *Number - Kind Code (if known)* Example: 1,234,567 B1	Publication Date MM-DD-YYYY	Name of Patentee or Applicant	Pages, Columns, Lines Where Relevant Passages or Relevant Figures Appear
DB	1	6,025,156	02-15-2000	Gwynn, et al.	
	2	6,124,433	09-26-2000	Falb, et al.	
	3	6,395,306 B1	05-28-2002	Cui, et al.	
	4	6,414,117 B1	07-02-2002	Levinson, D. A.	
DB		6,737,522 B2	05-18-2004	Sundick, et al.	

OIPE JC105 APR 13 2005 PATENT & TRADEMARK OFFICE

FOREIGN PATENT DOCUMENTS

Examiner Initials	Cite No.	Foreign Patent Document *Country Code-Number-Kind Code* Example: JP 1234567 A1	Publication Date MM-DD-YYYY	Name of Patentee or Applicant	Pages, Columns, Lines Where Relevant Passages or Relevant Figures Appear	T[1]

NON PATENT LITERATURE DOCUMENTS

Examiner Initials	Cite No.	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T[1]
DB	6	ALBERTS, et al. 1994. *Molecular Biology of the Cell, 3rd Edition*, pp. 403-404, 453. New York: Garland Publishing.	

1556093_1:dmb
040805

Examiner Signature	Date Considered 6/6/05

*Examiner: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

T[1] - Place a check mark in this area when an English language Translation is attached.

Notice of References Cited	Application/Control No. 10/063,557	Applicant(s)/Patent Under Reexamination EATON ET AL.	
	Examiner David J. Blanchard	Art Unit 1642	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
	A	US-6,277,574 B1	08-2001	Walker et al.	—
	B	US-			
	C	US-			
	D	US-			
	E	US-			
	F	US-			
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Hanish S. Nature Reviews, Applied Proteomics Collection, pp. 9-14, March 2005.
	V	Hanash et al. The Pharmacogenomics Journal, 3(6):308-311, 2003.
	W	Gygi et al. Molecular and Cellular Biology, 19(3):1720-1730, March 1999.
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

OPINION

Integrated global profiling of cancer

Samir Hanash

Tumours are complex biological systems. No single type of molecular approach fully elucidates tumour behaviour, necessitating analysis at multiple levels encompassing genomics and proteomics. Integrated data sets are required to fully determine the contributions of genome alterations, host factors and environmental exposures to tumour growth and progression, as well as the consequences of interactions between malignant or premalignant cells and their microenvironment. The sheer amount and heterogeneous nature of data that need to be collected and integrated are daunting, but effort has already begun to address these obstacles.

First published in *Nature Reviews Cancer* 4, 638–644 (2004)
doi:10.1038/nrc1409

In the 1980s, at the dawn of the era of molecular medicine, researchers believed that cancer was caused by dysregulation of a few oncogenes or tumour-suppressor genes. The identification of these genes would therefore lead to effective approaches for preventing or treating cancer. Substantial progress has been made in uncovering cancer genes that are altered through point mutations, deletions, amplifications, rearrangements or other events, and as a result effective targeted therapies for certain cancers have been developed. It has become clear, however, that human tumours are more complex and heterogeneous than expected, and are caused by defects in numerous pathways and factors that operate at many levels. For example, a gene can be amplified 100-fold in certain tumours with no demonstrable effect on RNA levels for that gene. Alternatively, protein levels can be increased, decreased or modified with no demonstrable changes in the levels of their corresponding RNAs. It is therefore a challenge to fully understand tumour behaviour, based on a single type of analysis. The factors that determine the consequences of a particular event or alteration can be highly context dependent, and are governed by the spatial and temporal activity of numerous interacting components. The intricate nature of the contributions of many factors ultimately determines the impact that a particular alteration has on the properties of a tumour or a precursor lesion.

There are two basic approaches to address the complexity of cancer. One is to reduce complexity through analysis of experimental models, such as cell lines or animal models, to characterize the fundamental processes of tumour growth and to elucidate the effects of single genes. Another is to integrate large data sets, to yield a model for tumour develop-ment and behaviour. Each approach has its own advantages and disadvantages. The first approach has been effective in many respects; for example, the early stages of tumorigenesis have been investigated using mouse models, and transformation and metastasis have been modelled in *Drosophila*[1]. However, in studying animal models of cancer, many factors that are relevant to human cancer are lost. The conclusions reached from these models are therefore not always applicable to human tumours[2]. The second approach, involving integration of large data sets, is challenging in part because only a limited number of samples, such as tumours or preneoplastic tissues, can be analysed in a given study. This makes data interpretation and model development difficult, given the large amount of heterogeneity between human tumours.

Profiling strategies

Improving our understanding of cancer and developing theoretical models will require an increased understanding of the contributions of and interactions between the numerous components that contribute to tumour formation and progression (FIG. 1). Strategies are available to profile changes at various levels, including the genome, transcriptome and proteome (TABLE 1). The host genome can be scanned for inherited variations such as mutations and polymorphisms that might contribute to cancer risk. Tumour cells and their precursors can be assayed for genomic alterations, such as chromosomal deletions or amplification, or changes in DNA methylation status, that promote their proliferation and survival. The cancer-cell transcriptome can be examined for patterns of gene expression, or its proteome analysed to uncover alterations in proteins, that contribute to tumour development or progression and would not be predicted by genome or transcriptome analysis.

A challenge for global profiling is the need to capture all the elements of the individual compartments that are profiled, such as the whole transcriptome or the whole proteome. Although this is possible for the transcriptome, other compartments, such as the proteome and metabolome, have numerous features that are difficult to capture, requiring several different profiling approaches (TABLE 1). For example, it is not possible to assay for protein functional activity, profile protein–protein interactions, and assess protein modifications all with the same platform. In all, there remains a substantial need to improve the breadth, sensitivity and throughput of global-profiling technologies.

In addition to global profiling of DNA, RNA or protein in normal, premalignant and malignant tissues, and in biological fluids, a comprehensive analysis would

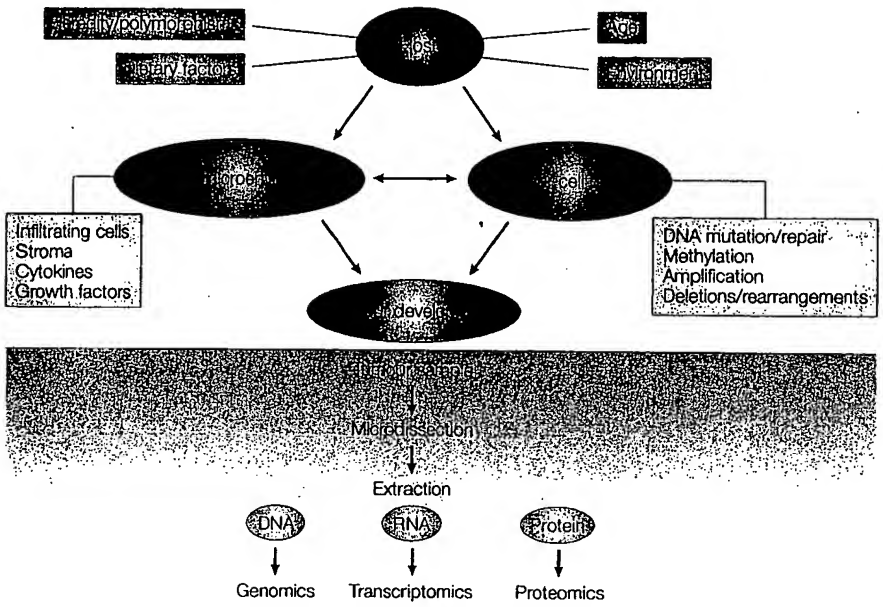


Figure 1 | **Numerous components must be integrated to study the molecular basis of human cancer.** Several host factors contribute to tumorigenesis in humans, including diet, environmental factors, polymorphisms and mutations in susceptibility genes, age and immunity. Cells undergo genomic changes (DNA mutations and repair, methylation, amplification, deletions and rearrangements), leading to tumorigenesis. Tumour development also depends on factors in the microenvironment — some of these are produced locally, whereas others are produced systemically (growth factors, infiltrating cells and cytokines). Reciprocal interactions between the premalignant and malignant cells, stromal cells, extracellular-matrix components, various inflammatory cells and a range of soluble mediators therefore contribute to tumour development and progression. Once tumour samples are obtained, genomic, transcriptomic and proteomic tools can be used to profile specific compartments.

me sure other characteristics from these samples to detect changes in nutritional, metabolic and immune status, as well as to detect environmental exposures. These types of data come from metabolic and nutritional profiles, immunohistochemical assays, assays of host immunity to tumour antigens, and patient questionnaires. Such data need to be integrated with molecular profile data.

Integrating data sets

So far, very few cancer studies have attempted to integrate data sets that were obtained by several different profiling techniques. Rather, the few large-scale integrated molecular-profiling efforts undertaken have combined data of a similar nature, notably combining transcriptome data obtained from several sources. Some studies have combined data obtained through two different global-profiling platforms (genomic and transcriptomic, or transcriptomic and proteomic) for the same set of study samples (such as lung tumours). These integrated data sets have also included variables such as clinical and pathological characteristics of the study individuals and their tumours, or mutations in cancer genes such as *TP53* and *RAS*. However limited in scope, these studies illustrate the potential impact of integrating data across numerous data sets in elucidating certain features of cancer[3–8].

Integrating gene-expression data from different sources. Profiling gene expression using DNA arrays has had a tremendous impact on biomedical research. Although the field is still in its infancy, there is increasing emphasis on integration of diverse sets of data. From a cancer research point of view, applications of global profiling of gene expression include uncovering unsuspected associations between genes, or identifying specific clinical features of cancer that result in novel molecular-based disease classifications. For example, DNA microarray analysis has been used to associate specific gene-expression profiles with different clinical outcomes of patients with the same types of tumours (responders versus non-responders[9]), or with cancer subtypes of the same lineage (high-stage versus low-stage tumours). Specific gene-expression signatures have also been associated with tumours of different lineages[10].

Lamb *et al.*[3] performed a study that illustrates the merits of integrating gene-expression data from several sources to develop a mechanistic understanding. They integrated gene-expression data from cell lines and human tumours to uncover a cyclin-dependent kinase (CDK)-independent mechanism of cyclin D1 function. Cyclin D1, which activates CDK, is frequently overexpressed in human tumours,

Table 1 | **Profiling strategies for genome-related components**

Platform	What we can learn	What is detected	Tools used for analysis
Genome	The hereditary components to cancer, as well as genome alterations in somatic cells that lead to cancer	Chromosome structural changes; gene copy-number changes; gene rearrangements; mutations/polymorphisms; methylation changes	DNA sequencing; cytogenetics; CGH; array CGH; SNP analysis; RLGS
Transcriptome	Changes in gene expression that are associated with cancer	Changes in RNA abundance; alterations in alternative splicing	Differential-display analysis; SAGE; DNA microarray analysis; PCR- and non-PCR-based gene-expression assays
Proteome	How proteins are modified or how their levels change in tumours	Protein levels; post-translational modifications; localization; protein–protein interactions; enzymatic activity	Sample-enrichment strategies (fractionation, protein tagging); separation-based profiling (2D gels, MS, LC, LC-MS); non-separation-based strategies (protein microarrays, direct MS analysis); protein-detection strategies (immunohistochemistry, immunofluorescence)

2D, two dimensional; CGH, comparative genomic hybridization; LC, liquid chromatography; MS, mass spectrometry; PCR, polymerase chain reaction; RLGS, restriction landmark genome scanning; SAGE, serial analysis of gene expression; SNP, single nucleotide polymorphism.

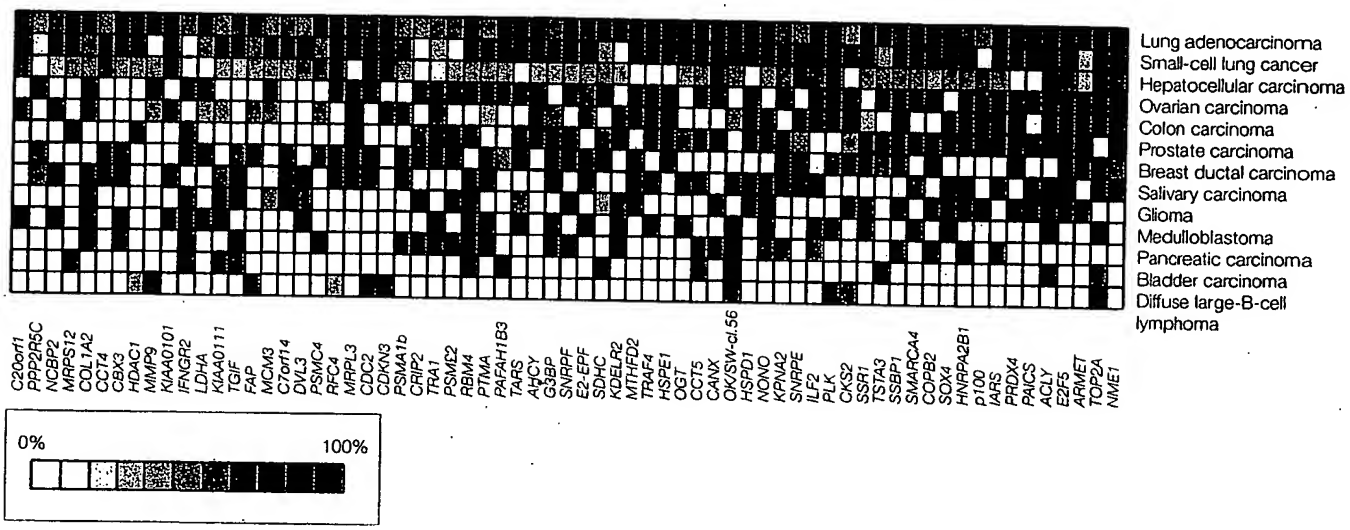


Figure 2 | **Integrated gene-expression profile of neoplastic transformation.** Public sharing of gene-expression data has led to the identification of 67 genes that are commonly overexpressed in tumour samples, relative to normal tissue. This 'meta-signature' analysis compared 'cancer versus normal' gene-expression signatures from 21 independent microarray data sets. Thirteen distinct cancer types were selected for this figure (listed on the right). White boxes signify genes for which no changes in expression were observed between tumour and normal cells. Light and dark red boxes signify genes that were significantly overexpressed in tumour cells, relative to normal tissue. Dark red indicates that the expression level was in the 90th percentile of all samples tested. Figure reproduced with permission from REF. 11 © (2004) National Academy of Sciences.

but the mechanisms by which this promotes tumorigenesis has been unclear. Cyclin D1 and a cyclin-D1 mutant that was incapable of activating CDK4 were each ectopically expressed in cultured human mammary epithelial cells. Twenty-one genes were found to be induced by both wild-type and mutant cyclin D1, indicating that these genes are CDK4 independent. Furthermore, the rapidity with which expression of these genes was induced indicated the direct involvement of a transcription factor. A database of gene-expression profiles from 190 primary human tumours was therefore also analysed, to identify cyclin-D1 target genes. The expression pattern of the set of 21 genes uncovered from *in vitro* studies was correlated with the levels of cyclin D1 in human tumours. A 'data-mining' process was applied to several human tumour gene-expression data sets, to identify genes that had a pattern of expression that matched the patterns of the genes that comprised the cyclin-D1 signature pattern. The transcription factor C/EBPβ was consistently co-expressed with the set of cyclin-D1 target genes. Functional analyses confirmed the involvement of C/EBPβ in the transcriptional regulation of cyclin D1. This study illustrates the types of findings that can be uncovered by integrating different sets of data.

Tumour gene-expression patterns are modulated by many extrinsic factors and by the microenvironment — these features could be crucial factors in determining the response to anticancer drugs. The gene-expression profiles of *in vitro* cultures of cancer cells have been compared with those of tumours grown *in vivo*, to determine the effects of the microenvironment on gene expression. In one study[4], two human cancer cell lines (a lung adenocarcinoma and a glioblastoma cell line) were transplanted into immunodeficient mice and allowed to form tumours, and the gene-expression profiles of these tumours were compared with those of cells grown in culture. A bioinformatics approach was used to associate genes into functional classes. The classes of genes that were expressed at higher levels in cells grown *in vitro* were associated with increased cell division and metabolism, reflecting the more favourable environment for cell proliferation. By contrast, *in vivo* tumour growth resulted in upregulation of a significant number of genes involved in extracellular-matrix formation, cell adhesion, cytokine and metalloproteinase activity, and neovascularization. When placed in comparable *in vivo* tissue environments, the lung cancer and the glioblastoma cells expressed different sets of extracellular-matrix- and cell-adhesion-related genes, indicating different mechanisms of extracellular interaction at work in the different tumour types. Importantly, gene products that are typically targeted by cancer therapies, such as tyrosine kinases, showed varied expression patterns when the same cancer cells were grown *in vitro* versus *in vivo*. This provides an indication of why therapeutics that are effective in *in vitro* studies might not always function *in vivo*.

A study that illustrates the merits of data sharing among investigators is a meta-analysis of cancer microarray data[11]. In this study, 40 published cancer microarray data sets comprising gene-expression measurements from over 3,700 tumour samples were collected and analysed. A common transcriptional profile that is activated in most cancer types, relative to corresponding normal tissues, was delineated from some of the data sets, providing a meta-signature of neoplastic transformation (FIG. 2).

Integrating genomic and transcriptomic data. Most tumours show numerous genomic alterations, but it has been a challenge to identify those that are required for different stages of tumour development. As most genome alterations — chromosomal gains and losses, deletions, amplification and methylation — affect the transcriptome, it would be useful to integrate genome profiling with transcriptome profiling. Several approaches are now available to scan the genome for gains and losses. These include fluorescence *in situ* hybridization, comparative genomic hybridization, hybridization of genomic DNA to various types of DNA microarrays, and restriction landmark genome scanning[5–8]. Additionally, oligonucleotide arrays are now available that can be used to detect single-nucleotide polymorphisms and that allow genome-wide loss-of-heterozygosity maps to be developed from tumours, including samples isolated by laser-capture microdissection[12].

Pollack *et al.* profiled DNA copy-number alterations across 6,691 mapped human genes in 44 samples of predominantly advanced, primary breast tumours and 10 breast cancer cell lines[13]. Parallel DNA microarray-based measurements of mRNA levels allowed assessment of the extent to which variation in gene copy number contributes to variation in gene expression in tumour cells. 62% of highly amplified genes showed increased expression levels. Additionally, DNA copy number correlated with gene expression across a range of DNA copy-number alterations, including deletions. On average, a twofold change in DNA copy number was associated with a corresponding 1.5-fold change in mRNA levels. It was estimated that overall, at least 12% of all the variation in gene expression among the breast tumours analysed was attributable to underlying variation in gene copy number, the remainder presumably attributable to a multitude of other factors.

In another study[14], restriction landmark genomic scanning was used to detect amplified genomic DNA fragments in 47 primary ovarian tumours. This approach uncovered amplification of the *LMYC* oncogene in several tumours. Transcriptome profiling of these tumours using oligonucleotide microarrays demonstrated frequent overexpression of *LMYC* in tumour cells, compared with cells of the normal ovarian surface epithelium — even in tumours without genomic amplification of *LMYC* — indicating that tumours use different mechanisms to upregulate *LMYC* expression. This finding prompted an assessment of the expression status of various members of the *MYC* gene family in ovarian tumours. Interestingly, a pattern was uncovered in which deregulated expression of one of the members of the *MYC* gene family was observed in most of the tumours.

Integrating transcriptome and proteome profiling. There is a need to profile gene expression at the level of the proteome and to correlate changes in gene-expression profiles with changes in proteomic profiles. The two are not always linked — numerous alterations occur in protein levels that are not reflected at the RNA level[15]. Translational control is an important cellular process that is regulated by several genes with tumour-suppressor or oncogenic properties[16]. For example, the proteins encoded by the tumour-suppressor genes tuberous sclerosis 1 (*TSC1*) and *TSC2* form a functional complex that inhibits the phosphorylation of S6 kinase and 4EBP1 — two key regulators of mRNA translation. TSC2 functions as a key regulator of the TOR pathway, which regulates protein synthesis, cell growth and viability in response to changes in cellular energy levels[17].

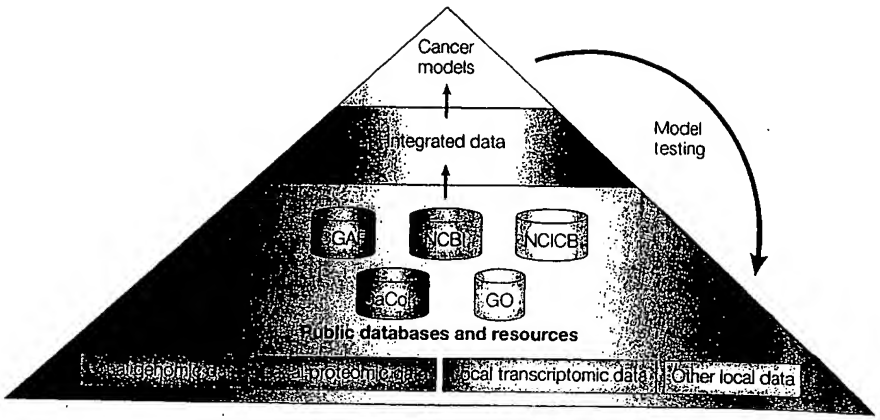


Figure 3 | **Path from data collection and integration to hypothesis testing.** Data produced by one research group can be combined with data in public databases such as the Cancer Genome Anatomy Project (CGAP) and further processed through resources available through various web sites — for example, the National Center for Biotechnology Information (NCBI), National Cancer Institute Center for Bioinformatics (NCICB), CaCore and Gene Ontology (GO) web sites — to yield integrated data sets (for further information on these web sites, see the online links box). This type of 'data mining' using statistical and informatics tools can lead to models for tumour behaviours such as metastasis, recurrence or response to therapy. Models can then be tested experimentally and/or through collection and analysis of additional data sets, and then refined.

Given the distinct regulation of RNA and protein levels, integration of data pertaining to RNA and protein products that are encoded by the same genes can tell us a lot about tumour function. Nishizuka *et al.* analysed gene-expression patterns of 60 human cancer cell lines (NCI-60) used by the National Cancer Institute to screen compounds for anticancer activity, and measured levels of 52 cancer-related proteins in these cells[18]. Clustered image maps of protein levels uncovered two markers that could be used to distinguish colon from ovarian adenocarcinomas. Integration of protein and mRNA data led to the interesting observation that the levels of structural proteins were highly correlated with the levels of their corresponding mRNAs in the NCI-60 cell lines, whereas the levels of non-structural proteins were poorly correlated with those of their corresponding mRNAs.

Gene-expression and proteomic data sets from lung tumours have also been compared and integrated, along with serum samples from the same patients[19–21]. To determine whether gene-expression profiles could be used in prognosis, mRNA profiles in tumours from 86 newly diagnosed patients, including 67 with early-stage and 19 with advanced-stage lung adenocarcinoma, were measured by oligonucleotide microarray analysis[19]. A gene-expression index, based on expression of the genes that correlated with survival of the 86 patients, was able to identify low-risk and high-risk groups among the patients with stage-I lung adenocarcinomas. The index included many novel genes that were not previously associated with survival in lung adenocarcinoma. A large number of genes, such as the *CRK* oncogene, showed a graded pattern of expression among the tumours. A small number of genes, such as *ERBB2*, were only overexpressed in a small number of tumours, but were also correlated with poor outcome.

In parallel, proteomic studies were undertaken to identify proteins associated with patient outcome[20]. A leave-one-out cross-validation procedure that analysed proteins associated with patient outcome — which were identified by Cox modelling — indicated that specific protein profiles can be used to predict the likelihood of survival in patients with stage-I tumours. Integration of RNA and protein data from the same tumours, and from an independent study, showed that 11 of 27 mRNAs associated with survival were represented in the profile of survival-associated proteins. Interestingly, combined analysis of protein and mRNA data revealed that 11 components of the glycolysis pathway were associated with poor outcome, either at the protein or RNA levels. Phosphoglycerate kinase 1 expression was associated with reduced patient survival time, based on both RNA and protein studies, and also based on immunohistochemistry analysis using tissue microarrays in an independent validation set of 117 lung tumours. The relative abundance of this protein in tumours led to the assessment of its levels in the sera of patients with lung cancer, revealing a correlation between increased serum levels of phosphoglycerate kinase 1 and poor outcome.

Challenges

The studies presented above, although relatively simple from the point of view of extent of integration of heterogeneous data sets, illustrate the merits of an integrated approach to tumour profiling. However, collecting and integrating sets of data that are quite diverse represents a substantial undertaking that necessitates resources not available to most investigators. Experimental data must be processed and stored in a manner that is compatible with integration with other external, scattered data sources. Further complications stem from the substantial variation in the nomenclature used to identify the same object and to designate its attributes. For example, the protein encoded by a gene can be designated differently from the gene itself. Annotation with controlled vocabularies is required to achieve comparability across data sets. Even with adequate resources, the data generated is not always sufficiently reliable for a meaningful integrated analysis. For example, for genes that are expressed at very low levels, mRNA and protein levels can show a lack of correlation simply because of the limited sensitivity of the measurements.

Another serious challenge to studying cancer pathogenesis is the effectiveness of developing models capable of accounting for all the data collected with different high-throughput approaches. Although researchers have attempted for many years to devise mathematical models for many aspects of cancer, such as for tumour growth[22], tumour drug delivery[23] or gene–environment interactions[24], it is challenging to develop models that integrate the numerous pathways and factors that operate at various levels during tumour growth. Development of a model that would be able to predict the consequences of a particular mutation for tumorigenesis is more difficult than predicting the consequences of a mutation for a simple system, such as for a cultured microorganism.

Models of human cancer are also impaired by the substantial lack of homogeneity among study populations and, most importantly, by the inability to manipulate components of the system. Furthermore, numerous members of the 'parts list' that is required to construct any model can not be measured or manipulated, or might not even have been identified. Initially, models might therefore represent approximations that generate hypotheses to be tested through further experimentation. Further experiments could then yield additional data to allow a more robust model to be developed (FIG. 3). For example, the finding that upregulation of glycolytic enzymes correlates with poor outcome in patients with lung cancer led to the finding that increased activity of the transcription factor hypoxia-inducible factor-1α (HIF1α), which is known to regulate expression of glycolytic-pathway genes, was also correlated with poor survival in patients with lung cancer[25]. HIF1α has since been associated with numerous tumour types.

Resources

An expanding array of resources, in the form of databases and tools, is available to allow experimental global profiling data and other types of data to be integrated. Fortunately, a large amount of data has become available on gene expression in normal and cancer cells through initiatives such as the Cancer Genome Anatomy Project and the Director's Challenge initiative, funded by the National Cancer Institute (NCI). There are also numerous other relevant data repositories. So an investigator who finds that a specific gene is upregulated in a certain tumour type would be able to learn more about the expression pattern of this gene in other tumour types, as well as in normal tissues, through various gene-expression databases (BOX 1).

There are now numerous resources available for mining data from various global-profiling techniques. One of the first publicly available web databases of pathway information is the Kyoto Encyclopedia of Genes and Genomes (KEGG)[26]. Over 150 pathways are represented with emphasis on well-defined metabolic pathways. The KEGG pathway reference diagrams can be readily integrated with genomic and proteomic data. GenMAPP (Gene MicroArray Pathway Profiler) is a freely available program for viewing and analyzing expression data on 'microarray pathway profiles' (MAPPs) representing biological pathways or any other functional grouping of genes[27]. Over 50 MAPP files depicting various biological pathways and gene families are available. GenMAPP includes gene annotation information as described by the Gene Ontology (GO) Consortium[28]. The GenMAPP program identifies GO terms that seem to be over-represented in a data set, providing clues to relevant biological processes. Transpath is an online web database on signal transduction and gene-regulatory pathways that lists over 15,000 protein–protein interactions involving several thousand genes[29]. The Kinase Pathway Database[30] uses a natural language processing algorithm to automatically extract protein interaction information from the literature.

Other resources include public databases of protein–protein interactions, namely the Biomolecular Interaction Database (BIND)[31] and the Database of Interacting Proteins[32]. However, the organism most represented in these databases is *Saccharomyces cerevisiae*, for which substantial protein–protein interaction data have been generated. (For further information on the resources discussed above and in the following section, see the online links box.)

Box 1 | Some of the resources available for 'data mining' in cancer research

In addition to maintaining the GenBank nucleic-acid sequence database, the National Center for Biotechnology Information (NCBI) provides data analysis and retrieval resources for the data in GenBank and other biological data made available through the NCBI web site[35]. Relevant NCBI resources include the Cancer Chromosome Aberration Project, Entrez Genomes and related tools, the Map Viewer, Model Maker, Evidence Viewer, the Clusters of Orthologous Groups (COGs) database, SAGEmap, Gene Expression Omnibus (GEO) and the Molecular Modeling Database (MMDB). There are also available custom implementations of the BLAST program that are optimized to search specialized data sets. The National Cancer Institute, through its Center for Bioinformatics, provides informatics infrastructure support to advance translational cancer research. The centre provides open access to large and diverse data sets that result from NCI-funded initiatives. It also provides a resource that integrates such data with outside data and provides facilities for data management and distribution. The resource, designated CaCore, consists of a series of component technologies and services[36]. Enterprise Vocabulary Services provide controlled vocabulary, dictionary and thesaurus services. The Cancer Data Standards Repository provides a meta-data registry for common data elements. Cancer Bioinformatics Infrastructure Objects (caBIO) implements an object-oriented model of the biomedical domain and provides Java, Simple Object Access Protocol and HTTP-XML application programming interfaces.

Other resources include GoMiner, developed by Zeeberg *et al.*[37]. GoMiner is a resource package that organizes lists of genes, such as under- and overexpressed genes from a range of microarray experiments[28]. GoMiner provides quantitative and statistical output files and visualization graph structures. Genes that are displayed in GoMiner are linked to the main public bioinformatics resources. (For further information on the resources discussed above, see the online links box.)

Future directions

Clearly, additional resources are needed to facilitate integration of diverse data sets. The NCI plans to deploy an integrating biomedical informatics infrastructure called the Cancer Biomedical Informatics Grid (CaBIG), which will be developed in partnership with the cancer-research community. Around 50 cancer centres have joined this NCI-led project. The goals of CaBIG are to integrate data from diverse sources and to support interoperable analytic tools. The open-source, open-access grid will allow different research groups to search the expanding collection of cancer research data together with locally generated data. A similar and related effort is also underway in the United Kingdom, where the National Cancer Research Institute, which represents government, philanthropic and private-sector organizations that fund cancer research, has set up a unit to develop cancer research informatics. This will facilitate integration of data generated by laboratories across different organizations.

Apart from informatics considerations, tumour-profiling technologies would benefit from miniaturization of assays and increases in throughput and sensitivity, given the limited availability of tumour tissues. For example, the availability of proteome-scale capture agents would facilitate the use of microarrays in proteomic profiling, in a manner similar to transcriptome profiling. The availability of technologies for global profiling using formalin-fixed tissue would also be beneficial.

Understanding cancer as a complex disease, through systems-biology or systems-pathology approaches, requires teams of investigators from diverse fields such as biomedicine, chemistry, engineering, informatics and computational modelling. Soon, data obtained from molecular imaging studies might also be integrated. The continued development of sensitive molecular-imaging-based assays that do not require tissue samples will be valuable for monitoring molecular and cellular processes in both animal models of cancer and in humans[33]. Integration of molecular imaging with other molecular approaches to tissue analysis could add a spatial and a temporal perspective to our understanding of tumour development and progression.

The need for multidisciplinary research into cancer and other diseases has been recognized by the National Institutes of Health (NIH) with the implementation of the 'NIH roadmap'[34]. A systems-biology approach to cancer that incorporates different genome-scale global-profiling technologies is expected to lead to the development of computational models of gene regulation in cancer and important cancer-related cell processes, such as differentiation, proliferation, transformation and metastasis. This will lead to molecular-based classifications of cancer that transcend organ and tissue types — these should supercede classifications based on histopathology or based on the expression patterns of genes with unknown functional significance. New and important features of tumorigenesis and tumour progression will be uncovered in this manner, leading to more effective screening strategies and therapeutic targets.

Samir Hanash is at the Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M5-C800, PO Box 19024, Seattle, Washington 98109, USA.
e-mail: shanash@fhcrc.org

1. Tapon, N. Modeling transformation and metastasis in *Drosophila*. *Cancer Cell* **4**, 333–335 (2003).
2. Rangarajan, A. & Weinberg, R. A. Comparative biology of mouse versus human cells: modelling human cancer in mice. *Nature Rev. Cancer* **3**, 952–359 (2003).
3. Lamb, J. *et al.* A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell* **114**, 323–334 (2003).
4. Creighton, C. *et al.* Profiling of pathway-specific changes in gene expression following growth of human cancer cell lines transplanted into mice. *Genome Biol.* **4**, R46 (2003).
5. Albertson, D. G., Collins, C., McCormick, F. & Gray, J. W. Chromosome aberrations in solid tumors. *Nature Genet.* **34**, 369–376 (2003).
6. Albertson, D. G. & Pinkel, D. Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.* **2**, R145–R152 (2003).
7. Shi, H. *et al.* Triple analysis of the cancer epigenome: an integrated microarray system for assessing gene expression, DNA methylation, and histone acetylation. *Cancer Res.* **63**, 2164–2171 (2003).
8. Feltus, F. A., Lee, E. K., Costello, J. F., Plass, C. & Vertino, P. M. Predicting aberrant CpG island methylation. *Proc. Natl Acad. Sci. USA* **100**, 12253–12258 (2003).
9. Ntzani, E. E. & Ioannidis, J. P. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* **362**, 1439–1444 (2003).
10. Simon, R. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Br. J. Cancer* **89**, 1599–1604 (2003).
11. Rhodes, D. R. *et al.* Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl Acad. Sci. USA* **101**, 9309–9314 (2004).
12. Lieberfarb, M. E. & Lin, M. Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP. *Cancer Res.* **63**, 4781–4785 (2003).
13. Pollack, J. R. *et al.* Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA* **99**, 12963–12968 (2002).
14. Wu, R. *et al.* Amplification and overexpression of the *L-MYC* proto-oncogene in ovarian carcinomas. *Am. J. Pathol.* **162**, 1603–1610 (2003).
15. Hanash, S. Disease proteomics. *Nature* **422**, 226–232 (2003).
16. Ruggero, D. & Pandolfi, P. P. Does the ribosome translate cancer? *Nature Rev. Cancer* **3**, 179–192 (2003).
17. Inoki, K., Zhu, T. & Guan, K. L. TSC2 mediates cellular energy response to control cell growth and survival. *Cell* **115**, 577–590 (2003).
18. Nishizuka, S. & Charboneau, L. Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proc. Natl Acad. Sci. USA* **100**, 14229–14234 (2003).
19. Beer, D. G. *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinomas. *Nature Med.* **8**, 816–824 (2002).
20. Chen, G. *et al.* Protein profiles associated with survival in lung adenocarcinoma. *Proc. Natl Acad. Sci. USA* **100**, 13537–13542 (2003).
21. Brichory, F. M. *et al.* An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer. *Proc. Natl Acad. Sci. USA* **98**, 9824–9829 (2001).
22. Albert, P. S. & Shih, J. H. Modeling tumor growth with random onset. *Biometrics* **59**, 897–906 (2003).
23. Telford, J. J., Saltzman, J. R., Kuntz, K. M. & Syngal, S. Impact of preoperative staging and chemoradiation versus postoperative chemoradiation on outcome in patients with rectal cancer: a decision analysis. *J. Natl Cancer Inst.* **96**, 191–201 (2004).
24. Merlino, G. & Noonan, F. P. Modeling gene-environment interactions in malignant melanoma. *Trends Mol. Med.* **9**, 102–108 (2003).
25. Semenza, G. L. Targeting HIF-1 for cancer therapy. *Nature Rev. Cancer* **3**, 721–732 (2003).
26. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
27. Doniger, S. W. *et al.* MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* **4**, R7 (2003).
28. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
29. Schacherer, F. *et al.* The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics* **17**, 1053–1057 (2001).
30. Koike, A., Kobayashi, Y. & Takagi, T. Kinase pathway database: An integrated protein-kinase and NLP-Based protein-interaction resources. *Genome Res.* **13**, 1231–1243 (2003).
31. Bader, G. D., Betel, D. & Hogue, C. W. BIND: The biomolecular interation network database. *Nucleic Acids Res.* **31**, 248–250 (2003).
32. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305 (2002).
33. Gelovani-Tjuvajev, J. & Blasberg, R. *In vivo* imaging of molecular-genetic targets for cancer therapy. *Cancer Cell* **4**, 327–333 (2003).
34. Zerhouni, E. The NIH Roadmap. *Science* **302**, 63–72 (2003).
35. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* **32** (Database issue), D35–D40 (2004).
36. Covitz, P. A. *et al.* caCORE: A common infrastructure for cancer informatics. *Bioinformatics* **19**, 2404–2412 (2003).
37. Zeeberg, B. R. *et al.* GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4**, R28 (2003).

Competing interests statement
The author declares no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:
Cancer.gov: http://cancer.gov/
breast cancer | lung cancer | ovarian cancer
Entrez Gene:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
CDK4 | *CRK* | cyclin D1 | *ERBB2* | *LMYC* | phosphoglycerate kinase 1 | *TP53* | *TSC1* | *TSC2*

FURTHER INFORMATION

ArrayExpress: http://www.ebi.ac.uk/arrayexpress
Biocarta: http://www.biocarta.com
Biomolecular Interaction Database:
http://www.blueprint.org/bind/bind.php
CaCORE: http://ncicb.nci.nih.gov/core
Cancer Biomedical Informatics Grid:
http://cabig.nci.nih.gov
Cancer Genome Anatomy Project: http://cgap.nci.nih.gov/
Cytoscape: http://www.cytoscape.org
Database of Interacting Proteins:
http://dip.doe-mbi.ucla.edu/

The Pharmacogenomics Journal (2003) 3, 308–311
© 2003 Nature Publishing Group All rights reserved 1470-269X/03 $25.00
www.nature.com/tpj

CLINICAL IMPLICATION

Making sense of microarray data to classify cancer

S Hanash[1] and C Creighton[2]

[1]*Department of Pediatrics, University of Michigan, Ann Arbor, MI, USA;* [2]*Bioinformatics Program, Ann Arbor, MI, USA*

The Pharmacogenomics Journal (2003) 3, 308–311. doi:10.1038/sj.tpj.6500209
Published online 4 November 2003

Profiling gene expression using DNA arrays has had a tremendous impact on biomedical research. From a disease investigation point of view, applications of DNA microarrays include uncovering unsuspected associations between genes and specific clinical features of disease, resulting in novel, molecular-based disease classifications. Cancer is a case in point. Most published studies of cancers using DNA microarrays have either examined a pathologically homogeneous set of tumors to identify clinically relevant subtypes, for example, responders *vs* nonresponders, or pathologically distinct subtypes of cancer of the same lineage, for example, high-stage *vs* low-stage tumors to identify molecular correlates, or tumors of different lineages to identify molecular signatures for each lineage. A study of cutaneous T-cell lymphoma by Kari *et al*,[1] published recently, typifies both what one hopes to gain from disease investigations using DNA microarrays and the limitations of such studies.

Primary cutaneous lymphomas are a heterogeneous group of lymphomas of T- or B-cell origin that represent a relatively common type of lymphoma and their incidence appears to be increasing. The two predominant subtypes of cutaneous T-cell lymphomas are mycosis fungoides, a mostly indolent variety, and its leukemic counterpart the Sezary syndrome, an aggressive variety characterized by skin involvement, lymphadenopathy and circulating atypical lymphocytes, the so-called Sezary cells. Kari *et al* used cDNA microarrays to study gene expression patterns in peripheral blood mononuclear cells from patients with the leukemic form of cutaneous T-cell lymphoma. The goal of the study was to identify markers that may be useful for diagnosis or prognosis, or that might provide new targets for treating this disease. The approach was to uncover gene expression differences between cells from 18 patients with high Sezary cell counts and an appropriate (Th2-skewed) cell fraction from nine normal controls. The differences in gene expression observed reflected many of the observed characteristics of the disease. Overexpressed genes in disease samples included some genes required for Th2 differentiation characteristic of Sezary cells. The analysis, however, did not uncover changes consistent with the hypothesis of defective apoptotic pathways in this disease. An important objective of the study was to identify markers for cutaneous T-cell lymphoma given the paucity of such markers. A member of the plastin gene family and a chemokine (CX3CR1) inappropriately expressed represented such potential novel markers. Two genes found to have a high predictive power to classify patients and controls were STAT4 and GTPase RhoB. These two genes alone accurately classified the high Sezary cell patients and controls. A signature profile with 10 genes was uncovered that identified a class of patients who succumb to the disease early, irrespective of their tumor burden. The study therefore uncovered a wealth of findings that shed some light on the biology of this disease and uncovered markers that may have a practical utility.

The DNA microarray studies described above and others in the literature indeed point to the great utility of DNA microarrays for uncovering patterns of gene expression that are clinically informative. Have the data been thoroughly analyzed? There is no shortage of analytical tools for uncovering patterns in microarray data. An important challenge for microarray analysis is to understand at a mechanistic level the significance of associations observed between subsets of genes and clinical features of disease. Another challenge is to identify the smallest but most informative sets of genes associated with specific clinical features, which then could be interrogated using technologies available in clinical laboratories, as appears to have been accomplished in this study. Another challenge is to determine how well RNA levels of predictive genes correlate with protein levels. A lack of correlation may imply that the predictive property of the gene(s) is independent of gene function.

To increase the effectiveness of DNA microarray analysis, global gene expression data may be combined with external data sources, such as gene annotation, in order to associate the expression patterns of a set of genes with the biological processes that they may represent. A welcome trend of data sharing allows others to analyze previously published microarray data and to combine multiple data sets. For illustration, we examined the data set published by Kari *et al* to see what we could uncover. In our analysis, we relied on the Gene Ontology (GO) annotation. The Gene Ontology Consortium[2] has defined a controlled vocabulary for describing genes in terms of their molecular function,

participation in biological processes and cellular locations. The GO annotations are making possible the high-throughput analyses of gene expression in terms of functional gene class associations, which otherwise would require laborious and somewhat subjective manual literature searches.

Using the data set from Kari *et al*, we searched a set of 122 genes found overexpressed in patients with high blood tumor burden, or Sezary cell count, compared to healthy controls ($P<0.01$, fold change >1.5), for significantly enriched (over-represented) GO terms, as described elsewhere.[3] We made the same search for a set of 280 genes found underexpressed in patients with high Sezary cell count ($P<0.01$, fold change <0.67). Our premise is that annotation terms that are shared by a significant number of genes within a large gene set may provide clues as to the processes driving the coordinate expression of the genes as a whole. Numerous enriched terms were found for the set of 280 underexpressed genes with $P<0.001$, including *class II major histocompatibility complex antigen* (five genes represented), *cytokine-binding activity* (six), *mitochondrion* (26), *electron transporter activity* (12) and *nucleotide metabolism* (four); these enriched terms could suggest a downregulation in CTCL of processes related to the immune response and mitochondrial function. Terms found enriched for the set of 122 overexpressed genes with $P<0.05$ include cell adhesion (nine genes represented) and cell cycle arrest (three).

The enriched GO terms listed above represent only a fraction of the genes significantly expressed in CTCL, and additional gene-to-process associations, not currently described in the biomedical literature or public annotation sources, may be inferred from data mining of large expression profile data sets. Our premise in this case is that genes that are coordinately expressed participate in closely related biological processes.[4] For a given gene, a GO term may be associated if the gene is correlated in expression with a significant number of other genes that share the given GO term annotation. We examined the expression patterns of 60 genes highly underexpressed in the Kari *et al* data set for patients with high Sezary cell count ($P<0.01$, fold change <0.33) that were also represented in a large independent data set of leukemia expression profiles from Armstrong *et al*.[5] For each of the 60 genes, the set of genes with significant positive correlations ($P<0.01$) with the given gene in the Armstrong data set was searched for significantly enriched GO terms ($P<0.0001$). In this way, 1963 gene-to-term associations, involving all 60 genes, were found. We performed two simulation tests to assess the number of random gene-to-term associations that could exist in the Armstrong data set, in one test permuting the expression values and

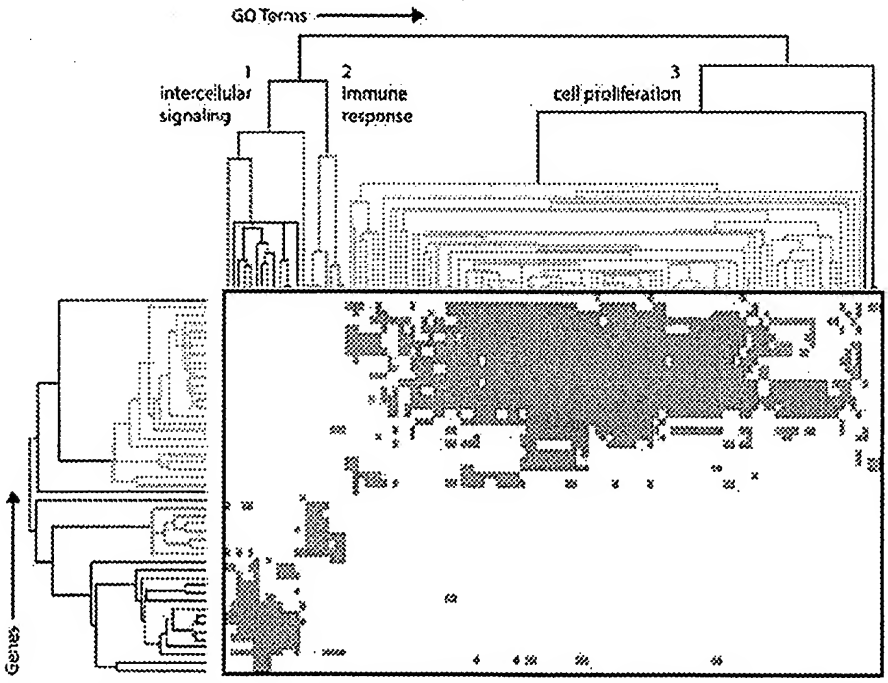


Figure 1 Hierarchical clustering of associations of GO terms for genes found underexpressed in patients with high Sezary cell count ($P<0.01$, fold change <0.33). For each gene-to-term association represented here, the given gene was found positively correlated in expression with a significant number of other genes that share the given GO term annotation. The rows in the matrix diagram represent genes; the columns represent terms. An entry in the matrix indicates that the corresponding gene-to-term association was found in the leukemia profile data set from Armstrong *et al* with $P<0.0001$. Three major clusters are highlighted corresponding to terms related to (1) intercellular signaling, (2) the immune response, and (3) cell proliferation. Table 1 lists the genes that fall under each cluster.

Table 1 GO term associations from Figure 1 for genes underexpressed in patients with high Sezary cell counts

Gene	Gene product description
Cluster 1—integral to plasma membrane; receptor activity; signal transducer activity; cell surface receptor-linked signal transduction; cell motility; G-protein-coupled receptor protein signaling pathway; cell–cell signaling; development; organogenesis; morphogenesis; extracellular	
CCL2	Small inducible cytokine A2
CD8B1	CD8 antigen, beta polypeptide 1 (p37)
CTSL	Cathepsin L
GPNMB	Glycoprotein (transmembrane) nmb
IL1R1	Interleukin 1 receptor, type I
ITGB4	Integrin, beta 4
MAL	Mal, T-cell differentiation protein
MAOA	Monoamine oxidase A
ME1	Malic enzyme 1, NADP(+)-dependent, cytosolic
PLAU	Plasminogen activator, urokinase
STAT4	Signal transducer and activator of transcription 4
TNFAIP6	Tumor necrosis factor, alpha-induced protein 6
Cluster 2—immune response; response to biotic stimulus; defense response; vacuole; lytic vacuole; lysosome	
CCL4	Small inducible cytokine A4 (homologous to mouse Mip-1b)
CYP1B1	Cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile)
FCER2	Fc fragment of IgE, low-affinity II, receptor for (CD23A)
GZMK	Granzyme K (serine protease, granzyme 3; tryptase II)
IL4R	Interleukin 4 receptor
MMP9	Matrix metalloproteinase 9 (gelatinase B, 92 kDa gelatinase, 92 kDa type IV collagenase)
TIMP1	Tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibitor)
Cluster 3—DNA repair; DNA replication; nucleolus; cell cycle; cell proliferation; mitosis; mRNA processing; mRNA splicing; ubiquitin-dependent protein catabolism; 26S proteasome; spliceosome complex; translation initiation factor activity; mitochondrion; oxidative phosphorylation; tricarboxylic acid cycle; cytochrome c oxidase activity	
AKAP9	A kinase (PRKA) anchor protein (yotiao) 9
AP1B1	Adaptor-related protein complex 1, beta 1 subunit
ATOX1	ATX1 (antioxidant protein 1, yeast) homolog 1
ATP5G3	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit c (subunit 9) isoform 3
CD164	CD164 antigen, sialomucin
CDC2	Cell division cycle 2, G1–S and G2–M
DRG1	Developmentally regulated GTP-binding protein 1
HADH2	Hydroxyacyl-coenzyme A dehydrogenase, type II
HLA-DQB1	Major histocompatibility complex, class II, DQ beta 1
LDHA	Lactate dehydrogenase A
LMNB2	Lamin B2
NDUFS1	NADH dehydrogenase (ubiquinone) Fe-S protein 1 (75 kDa) (NADH-coenzyme Q reductase)
OXCT	3-oxoacid CoA transferase
PCNA	Proliferating cell nuclear antigen
RUNX1	Runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene)
SATB1	Special AT-rich sequence-binding protein 1 (binds to nuclear matrix/scaffold-associating DNA's)
SLC25A11	Solute carrier family 25 (mitochondrial carrier; oxoglutarate carrier), member 11
SPINT2	Serine protease inhibitor, Kunitz type, 2
TOP2A	Topoisomerase (DNA) II alpha (170 kDa)
TXNRD1	Thioredoxin reductase 1
UBE2C	Ubiquitin carrier protein E2-C
VDAC1	Voltage-dependent anion channel 1
YWHAZ	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide
ZNF204	Zinc-finger protein 204

in another test permuting the annotation assignments. Neither search with these randomized data sets yielded more than 15 associations, indicating that most of the actual associations found were not the result of chance.

We used average linkage hierarchical clustering[4] to obtain a global view of the gene-to-term associations mined from the Armstrong leukemia expression data set. Figure 1 shows the resulting cluster diagram (with 113 GO terms that were associated with at least five genes being represented and with 49 genes that were associated with at least one of these terms). Genes are represented in the rows of the matrix diagram, and

terms are represented in the columns. An entry in the diagram indicates that the given gene (underexpressed in patients with high Sezary count compared to healthy controls) was coexpressed with a significant number of genes that share the given annotation term. GO terms that are closely related to each other by the biological concepts that they represent were found to cluster together. The clustering diagram defines three distinct major clusters of genes and terms related to intercellular signaling (labeled as Cluster '1' in the figure), the immune response (labeled Cluster '2'), and cell proliferation (labeled Cluster '3'). Table 1 lists the genes that fall under each cluster, with example terms.

Our GO term clustering analysis indicates that many of the genes underexpressed in CTCL may be associated with processes of cell proliferation, the immune response or intercellular signaling, which suggests a hypothesis that the pathogenesis of CTCL involves a downregulation of these processes. CTCL is characterized by the accumulation of malignant cells with a low proliferative index, which appears consistent with the observation made here of numerous genes associated with proliferation being underexpressed in CTCL. The observed underexpression in CTCL of numerous genes involved in the immune response, including several genes encoding for the class II major histocompatibility antigen complex, might be construed as contradicting one hypothesis that CTCL may be a malignancy of T cells stimulated to proliferate against its own tumor antigens.[6] There has been much speculation that CTCL cells are defective in their apoptotic pathways, and that the disease is linked to an accumulation rather than a true proliferation of T cells.[7] Underexpressed genes in CTCL thought to mediate apoptosis, including STAT4, CTSL (cathepsin L), IL1R1 (interleukin 1 receptor, type I) and TNFAIP6 (tumor necrosis factor, alpha-induced protein 6), are associated here with intercellular signaling-related terms.

However perfected DNA microarrays and their analytical tools become for disease profiling, they will not eliminate a pressing need for other types of profiling technologies that go beyond measuring RNA levels, particularly for disease-related investigations. DNA microarrays have limited utility for the analysis of biological fluids and for uncovering directly in the fluid, assayable biomarkers. There is a need to assay protein levels and activity. Numerous alterations may occur in proteins that are not reflected in changes at the RNA level, providing a compelling rationale for additional, direct analysis of gene expression at the protein level. The next challenge is to integrate RNA data with protein data.

DUALITY OF INTEREST

None declared

Correspondence should be sent to:
SM Hanash, Department of Pediatrics, University of Michigan, 1150 W. Medical Center Drive, MSRBI, Room A520, Ann Arbor, MI 48109, USA.
Tel: +734 763 9311
Fax: +734 647 8148
E-mail: shanash@umich.edu

REFERENCES

1 Kari L, Loboda A, Nebozhyn M, Rook AH, Vonderheid EC, Nichols C *et al.* Classification and prediction of survival in patients with the leukemic phase of cutaneous T cell lymphoma. *J Exp Med* 2003; **197**: 1477–1488.
2 Ashburner M, Ball CA, Blake JA, Botstein D, Cherry JM *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**: 25–29.
3 Creighton C, Kuick R, Misek DE, Rickman DS, Brichory FM, Rouillard JM *et al.* Profiling of pathway-specific changes in gene expression following growth of human cancer cell lines transplanted into mice. *Genome Biol* 2003; **4**: R46.
4 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998; **95**: 14863–14868.
5 Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* 2002; **30**: 41–47.
6 Edelson RL. Cutaneous T cell lymphoma: the helping hand of dendritic cells. *Ann NY Acad Sci* 2001; **941**: 1–11.
7 de Arruda MV, Watson S, Lin C-S, Leavitt J, Matsudaira P. Fimbrin is a homologue of the cytoplasmic phosphoprotein plastin and has domains homologous with calmodulin and actin gelation proteins. *J Cell Biol* 1990; **11**: 1069–1079.

MOLECULAR AND CELLULAR BIOLOGY, Mar. 1999, p. 1720–1730 Vol. 19, No. 3
0270-7306/99/$04.00+0
Copyright © 1999, American Society for Microbiology. All Rights Reserved.

Correlation between Protein and mRNA Abundance in Yeast

STEVEN P. GYGI, YVAN ROCHON, B. ROBERT FRANZA, AND RUEDI AEBERSOLD*
Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730

Received 5 October 1998/Returned for modification 11 November 1998/Accepted 2 December 1998

We have determined the relationship between mRNA and protein expression levels for selected genes expressed in the yeast *Saccharomyces cerevisiae* growing at mid-log phase. The proteins contained in total yeast cell lysate were separated by high-resolution two-dimensional (2D) gel electrophoresis. Over 150 protein spots were excised and identified by capillary liquid chromatography-tandem mass spectrometry (LC-MS/MS). Protein spots were quantified by metabolic labeling and scintillation counting. Corresponding mRNA levels were calculated from serial analysis of gene expression (SAGE) frequency tables (V. E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett, Jr., P. Hieter, B. Vogelstein, and K. W. Kinzler, Cell 88:243–251, 1997). We found that the correlation between mRNA and protein levels was insufficient to predict protein expression levels from quantitative mRNA data. Indeed, for some genes, while the mRNA levels were of the same value the protein levels varied by more than 20-fold. Conversely, invariant steady-state levels of certain proteins were observed with respective mRNA transcript levels that varied by as much as 30-fold. Another interesting observation is that codon bias is not a predictor of either protein or mRNA levels. Our results clearly delineate the technical boundaries of current approaches for quantitative analysis of protein expression and reveal that simple deduction from mRNA transcript analysis is insufficient.

The description of the state of a biological system by the quantitative measurement of the system constituents is an essential but largely unexplored area of biology. With recent technical advances including the development of differential display-PCR (21), of cDNA microarray and DNA chip technology (20, 27), and of serial analysis of gene expression (SAGE) (34, 35), it is now feasible to establish global and quantitative mRNA expression profiles of cells and tissues in species for which the sequence of all the genes is known. However, there is emerging evidence which suggests that mRNA expression patterns are necessary but are by themselves insufficient for the quantitative description of biological systems. This evidence includes discoveries of posttranscriptional mechanisms controlling the protein translation rate (15), the half-lives of specific proteins or mRNAs (33), and the intracellular location and molecular association of the protein products of expressed genes (32).

Proteome analysis, defined as the analysis of the protein complement expressed by a genome (26), has been suggested as an approach to the quantitative description of the state of a biological system by the quantitative analysis of protein expression profiles (36). Proteome analysis is conceptually attractive because of its potential to determine properties of biological systems that are not apparent by DNA or mRNA sequence analysis alone. Such properties include the quantity of protein expression, the subcellular location, the state of modification, and the association with ligands, as well as the rate of change with time of such properties. In contrast to the genomes of a number of microorganisms (for a review, see reference 11) and the transcriptome of *Saccharomyces cerevisiae* (35), which have been entirely determined, no proteome map has been completed to date.

The most common implementation of proteome analysis is the combination of two-dimensional gel electrophoresis (2DE) (isoelectric focusing-sodium dodecyl sulfate [SDS]-polyacrylamide gel electrophoresis) for the separation and quantitation of proteins with analytical methods for their identification. 2DE permits the separation, visualization, and quantitation of thousands of proteins reproducibly on a single gel (18, 24). By itself, 2DE is strictly a descriptive technique. The combination of 2DE with protein analytical techniques has added the possibility of establishing the identities of separated proteins (1, 2) and thus, in combination with quantitative mRNA analysis, of correlating quantitative protein and mRNA expression measurements of selected genes.

The recent introduction of mass spectrometric protein analysis techniques has dramatically enhanced the throughput and sensitivity of protein identification to a level which now permits the large-scale analysis of proteins separated by 2DE. The techniques have reached a level of sensitivity that permits the identification of essentially any protein that is detectable in the gels by conventional protein staining (9, 29). Current protein analytical technology is based on the mass spectrometric generation of peptide fragment patterns that are idiotypic for the sequence of a protein. Protein identity is established by correlating such fragment patterns with sequence databases (10, 22, 37). Sophisticated computer software (8) has automated the entire process such that proteins are routinely identified with no human interpretation of peptide fragment patterns.

In this study, we have analyzed the mRNA and protein levels of a group of genes expressed in exponentially growing cells of the yeast *S. cerevisiae*. Protein expression levels were quantified by metabolic labeling of the yeast proteins to a steady state, followed by 2DE and liquid scintillation counting of the selected, separated protein species. Separated proteins were identified by in-gel tryptic digestion of spots with subsequent analysis by microspray liquid chromatography-tandem mass spectrometry (LC-MS/MS) and sequence database searching. The corresponding mRNA transcript levels were calculated from SAGE frequency tables (35).

This study, for the first time, explores a quantitative comparison of mRNA transcript and protein expression levels for a relatively large number of genes expressed in the same metabolic state. The resultant correlation is insufficient for predic-

* Corresponding author. Mailing address: Department of Molecular Biotechnology, Box 357730, University of Washington, Seattle, WA 98195-7730. Phone: (206) 221-4196. Fax: (206) 685-7301. E-mail: ruedi@u.washington.edu.

BEST AVAILABLE COPY

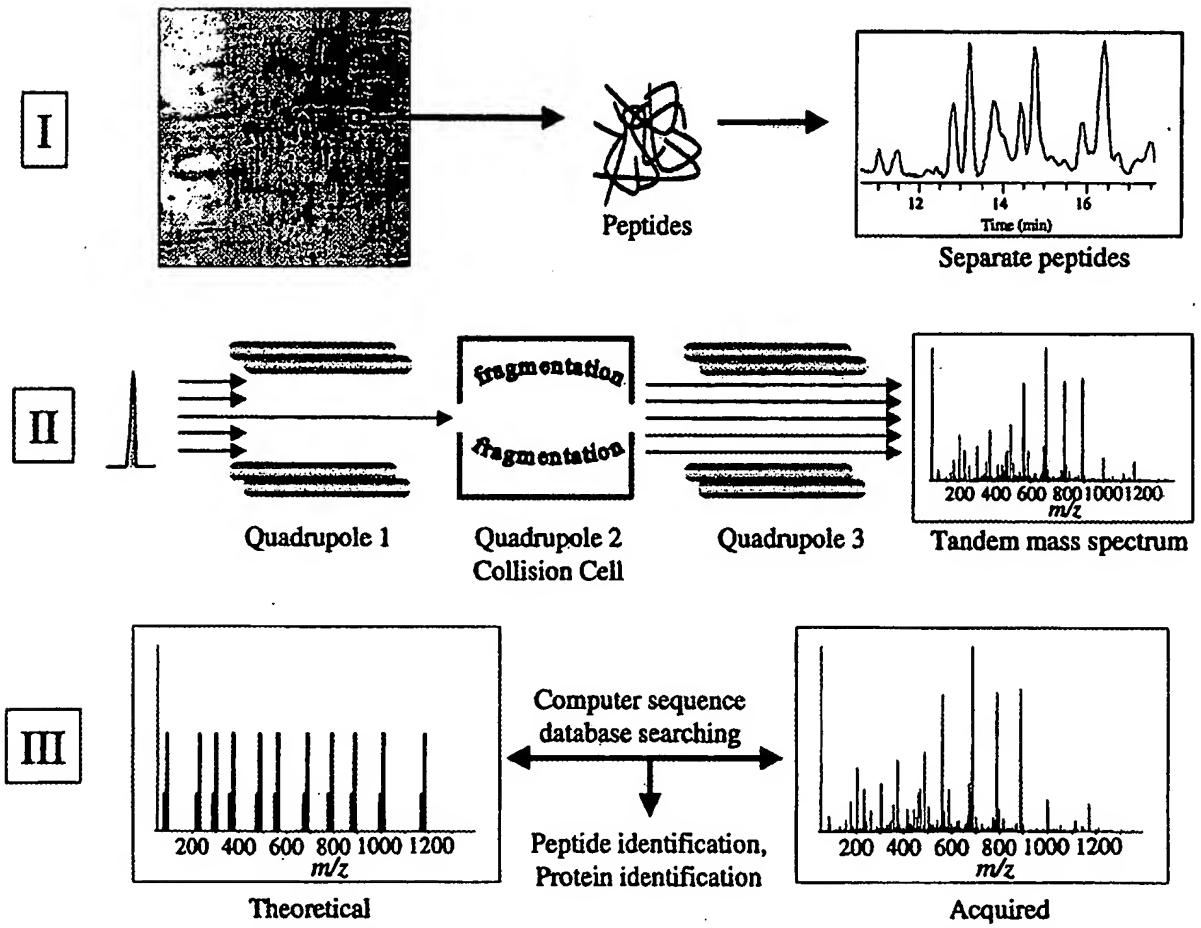


FIG. 1. Schematic illustration of proteome analysis by 2DE and mass spectrometry. In part I, proteins are separated by 2DE, stained spots are excised and subjected to in-gel digestion with trypsin, and the resulting peptides are separated by on-line capillary high-performance liquid chromatography. In part II, a peptide is shown eluting from the column in part I. The peptide is ionized by electrospray ionization and enters the mass spectrometer. The mass of the ionized peptide is detected, and the first quadrupole mass filter allows only the specific mass-to-charge ratio of the selected peptide ion to pass into the collision cell. In the collision cell, the energized, ionized peptides collide with neutral argon gas molecules. Fragmentation of the peptide is essentially random but occurs mainly at the peptide bonds, resulting in smaller peptides of differing lengths (masses). These peptide fragments are detected as a tandem mass (MS/MS) spectrum in the third quadrupole mass filter where two ion series are recorded simultaneously, one each from sequencing inward from the N and C termini of the peptide, respectively. In part III, the MS/MS spectrum from the selected, ionized peptide is compared to predicted tandem mass spectra computer generated from a sequence database. Provided that the peptide sequence exists in the database, the peptide and, by association, the protein from which the peptide was derived can be identified. Unambiguous protein identification is attained in a single analysis because multiple peptides are identified as being derived from the same protein.

tion of protein levels from mRNA transcript levels. We have also compared the relative amounts of protein and mRNA with the respective codon bias values for the corresponding genes. This comparison indicates that codon bias by itself is insufficient to accurately predict either the mRNA or the protein expression levels of a gene. In addition, the results demonstrate that only highly expressed proteins are detectable by 2DE separation of total cell lysates and that therefore the construction of complete proteome maps with current technology will be very challenging, irrespective of the type of organism.

MATERIALS AND METHODS

Yeast strain and growth conditions. The source of protein and message transcripts for all experiments was YPH499 (*MATa ura3-52 lys2-801 ade2-101 leu2-Δ1 his3-Δ200 trp1-Δ63*) (30). Logarithmically growing cells were obtained by growing yeast cells to early log phase (3×10^6 cells/ml) in YPD rich medium (YPD supplemented with 6 mM uracil, 4.8 mM adenine, and 24 mM tryptophan) at 30°C (35). Metabolic labeling of protein was accomplished in YPD medium exactly as described elsewhere (4) with the exception that 1 ml of cells was labeled with 3 mCi to offset methionine present in YPD medium. Protein was harvested as described by Garrels and coworkers (12). Harvested protein was lyophilized, resuspended in isoelectric focusing gel rehydration solution, and stored at −80°C.

2DE. Soluble proteins were run in the first dimension by using a commercial flatbed electrophoresis system (Multiphor II; Pharmacia Biotech). Immobilized polyacrylamide gel (IPG) dry strips with nonlinear pH 3.0 to 10.0 gradients (Amersham-Pharmacia Biotech) were used for the first-dimension separation. Forty micrograms of protein from whole-cell lysates was mixed with IPG strip rehydration buffer (8 M urea, 2% Nonidet P-40, 10 mM dithiothreitol), and 250 to 380 μl of solution was added to individual lanes of an IPG strip rehydration tray (Amersham-Pharmacia Biotech). The strips were allowed to rehydrate at room temperature for 1 h. The samples were run at 300 V–10 mA–5 W for 2 h, then ramped to 3,500 V–10 mA–5 W over a period of 3 h, and then kept at 3,500 V–10 mA–5 W for 15 to 19 h. At the end of the first-dimension run (60 to 70 kV · h), the IPG strips were reequilibrated for 8 min in 2% (wt/vol) dithiothreitol in 2% (wt/vol) SDS–6 M urea–30% (wt/vol) glycerol–0.05 M Tris HCl (pH 6.8) and for 4 min in 2.5% iodoacetamide in 2% (wt/vol) SDS–6 M urea–30% (wt/vol) glycerol–0.05 M Tris HCl (pH 6.8). Following reequilibration, the strips were transferred and apposed to 10% polyacrylamide second-dimension gels. Polyacrylamide gels were poured in a casting stand with 10% acrylamide–2.67% piperazine diacrylamide–0.375 M Tris base-HCl (pH 8.8)–0.1% (wt/vol) SDS–0.05%

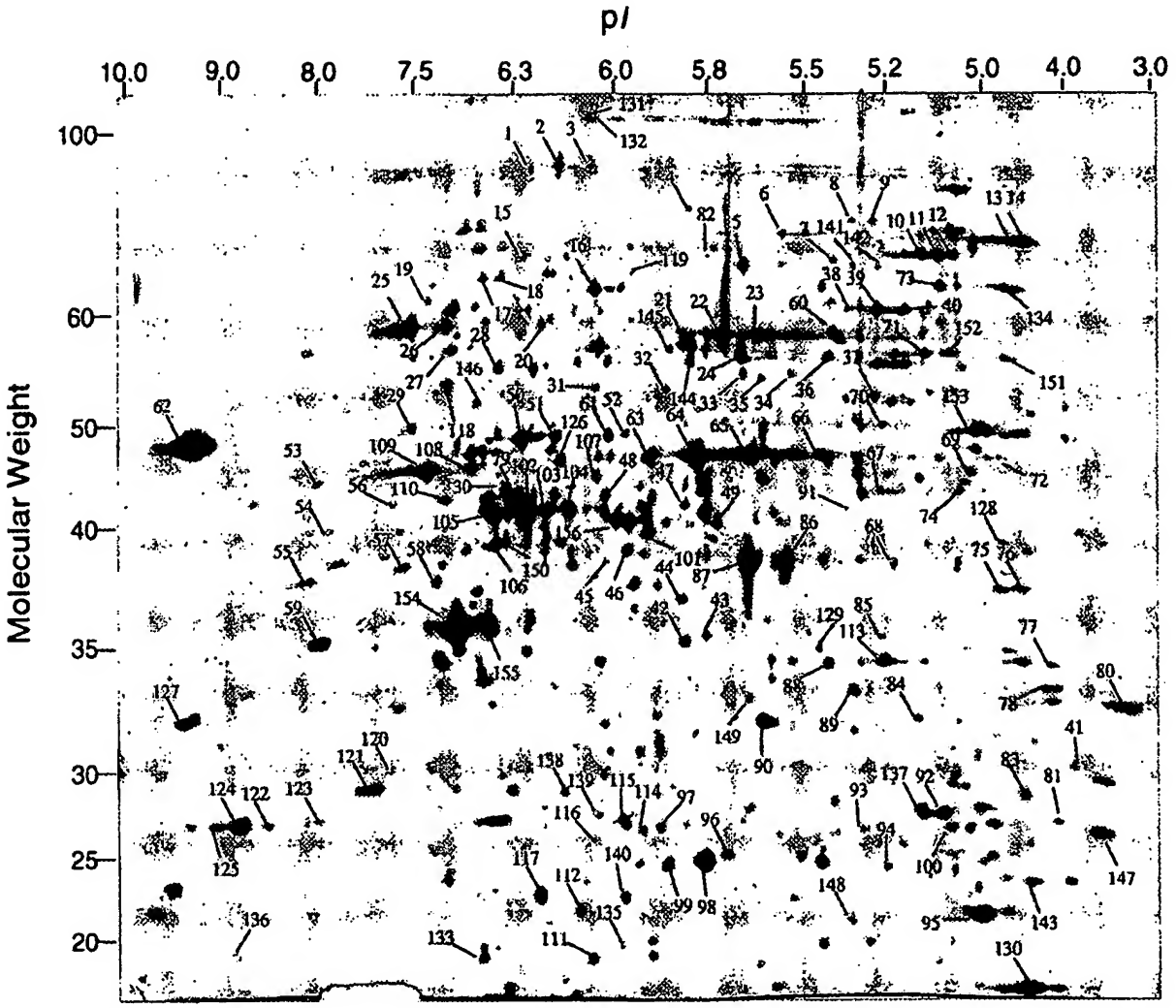


FIG. 2. 2D silver-stained gel of the proteins in yeast total cell lysate. Proteins were separated in the first dimension (horizontal) by isoelectric focusing and then in the second dimension (vertical) by molecular weight sieving. Protein spots (156) were chosen to include the entire range of molecular weights, isoelectric focusing points, and staining intensities. Spots were excised, and the corresponding protein was identified by mass spectrometry and database searching. The spots are labeled on the gel and correspond to the data presented in Table 1. Molecular weights are given in thousands.

(wt/vol) ammonium persulfate–0.05% TEMED (*N,N,N',N'*-tetramethylethylenediamine) in Milli-Q water. The apparatus used to run second-dimension gels was a noncommercial apparatus from Oxford Glycosciences, Inc. Once the IPG strips were apposed to the second-dimension gels, they were immediately run at 50 mA (constant)–500 V–85 W for 20 min, followed by 200 mA (constant)–500 V–85 W until the buffer front line was 10 to 15 mm from the bottom of the gel. Gels were removed and silver stained according to the procedure of Shevchenko et al. (29).

Protein identification. Gels were exposed to X-ray film overnight, and then the silver staining and film were used to excise 156 spots of varying intensities, molecular weights, and isoelectric focusing points. In order to increase the detection limit by mass spectrometry, spots were cut out and pooled from up to four identical cold, silver-stained gels. In-gel tryptic digests of pooled spots were performed as described previously (29). Tryptic peptides were analyzed by microcapillary LC-MS with automated switching to MS/MS mode for peptide fragmentation. Spectra were searched against the composite OWL protein sequence database (version 30.2; 250,514 protein sequences) (24a) by using the computer program Sequest (8), which matches theoretical and acquired tandem mass spectra. A protein match was determined by comparing the number of peptides identified and their respective cross-correlation scores. All protein identifications were verified by comparison with theoretical molecular weights and isoelectric points.

mRNA quantitation. Velculescu and coworkers have previously generated frequency tables for yeast mRNA transcripts from the same strain grown under the same stated conditions as described herein (35). The SAGE technology is based on two main principles. First, a short sequence tag (15 bp) that contains sufficient information uniquely to identify a transcript is generated. A single tag is usually generated from each mRNA transcript in the cell which corresponds to 15 bp at the 3'-most cutting site for *Nla*III. Second, many transcript tags can be concatenated into a single molecule and then sequenced, revealing the identity of multiple tags simultaneously. Over 20,000 transcripts were sequenced from yeast strain YPH499 growing at mid-log phase on glucose. Assuming the previously derived estimate of 15,000 mRNA molecules per cell (16), this would represent a 1.3-fold coverage even for mRNA molecules present at a single copy per cell and would provide a 72% probability of detecting such transcripts. Computer software which took for input the gene detected, examined the nucleotide sequence, and performed the calculation as described by Velculescu and coworkers (35) was written. In practice, we found that for 21 of 128 (16%) genes examined viable mRNA levels from SAGE data could not be calculated. This was because (i) no CATG site was found in the open reading frame (ORF), (ii) a CATG site was found but the corresponding 10-bp putative SAGE tag was not found in the frequency tables, or (iii) identical putative SAGE tags were present for multiple genes (e.g., TDH2_YEAST and TDH3_YEAST).

TABLE 1. Expressed genes identified from 2D gel in Fig. 2

Mol wt	pI	Spot no.	YPD gene name[a]	Protein abundance (10^3 copies/cell)	mRNA abundance (copies/cell)	Codon bias
17,259	6.75	133	CPR1	15.2	61.7	0.769
18,702	4.80	83	EGD2	20.1	5.2	0.724
18,726	4.44	147	YKL056C	61.2	88.4	0.831
18,978	5.95	135	YER067W	3.7	6.7	0.118
19,108	5.04	130	YLR109W	94.4	9.7	0.680
19,681	9.08	136	ATP7	11.0	NA[b,c]	0.246
20,505	6.07	111	GUK1	16.5	3.7	0.422
21,444	5.25	148	SAR1	5.4	10.4	0.455
21,583	4.98	95	TSA1	110.6	40.1	0.845
22,602	4.30	80	EFB1	66.1	23.8	0.875
23,079	6.29	112	SOD2	12.6	2.2	0.351
23,743	5.44	137	HSP26	NA[d]	0.7	0.434
24,033	5.97	96	ADK1	17.4	16.4	0.656
24,058	4.43	143	YKL117W	29.2	10.4	0.339
24,353	6.30	140	TFS1	8.1	0.7	0.146
24,662	5.85	99	URA5	25.4	6.0	0.359
24,808	6.33	97	GSP1	26.3	5.2	0.735
24,908	8.73	122	RPS5	18.6	NA[c]	0.899
25,081	4.65	81	MRP8	9.3	NA[c]	0.241
25,960	6.06	116	RPE1	5.8	0.7	0.372
26,378	9.55	127	RPS3	96.8	NA[c]	0.863
26,467	5.18	100	VMA4	10.5	3.7	0.427
26,661	5.84	98	TPI1	NA[d]	NA[c]	0.900
27,156	5.56	93	PRE8	6.9	0.7	0.129
27,334	6.13	115	YHR049W	18.4	2.2	0.520
27,472	5.33	92	YNL010W	31.6	3.7	0.421
27,480	8.95	123	GPM1	10.0	169.4	0.902
27,480	8.95	124	GPM1	231.4	169.4	0.902
27,480	8.95	125	GPM1	7.5	169.4	0.902
27,809	5.97	139	HOR2	5.7	0.7	0.381
27,874	4.46	78	YST1	13.6	52.8	0.805
28,595	4.51	41	PUP2	4.4	0.7	0.147
29,156	6.59	114	YMR226C	14.5	2.2	0.283
29,244	8.40	120	DPM1	5.0	11.2	0.362
29,443	5.91	48	PRE4	3.4	3.7	0.162
30,012	6.39	138	PRB1	21.2	1.5	0.449
30,073	4.63	77	BMH1	14.7	28.2	0.454
30,296	7.94	121	OMP2	67.4	41.6	0.499
30,435	6.34	89	GPP1	70.2	11.2	0.703
31,332	5.57	88	ILV6	13.9	3.0	0.402
32,159	5.46	113	IPP1	63.1	3.7	0.752
32,263	6.00	149	HIS1	22.4	4.5	0.232
33,311	5.35	84	SPE3	15.1	6.7	0.468
34,465	5.60	129	ADE1	8.7	5.2	0.305
34,762	5.32	85	SEC14	10.9	6.0	0.373
34,797	5.85	42	URA1	49.5	8.9	0.237
34,799	6.04	90	BEL1	103.2	81.0	0.875
35,556	5.97	43	YDL124W	6.4	4.5	0.206
35,619	8.41	59	TDH1	69.8	32.7[e]	0.940
35,650	5.49	68	CAR1	5.2	3.0	0.339
35,712	6.72	117	TDH2	49.6	473.0[e]	0.982
35,712	6.72	154	TDH2	863.5	473.0[e]	0.982
35,712	6.72	155	TDH2	79.4	473.0[e]	0.982
36,272	4.85	128	APA1	8.7	0.7	0.425
36,358	5.05	75	YJR105W	17.6	17.1	0.522
36,358	5.05	76	YJR105W	27.5	17.1	0.522
36,596	6.37	79	ADH2	58.9	260.0[e]	0.711
36,714	6.30	102	ADH1	746.1	260.0	0.913
36,714	6.30	103	ADH1	17.6	260.0	0.913
36,714	6.30	104	ADH1	61.4	260.0	0.913
36,714	6.30	105	ADH1	52.7	260.0	0.913
37,033	6.23	44	TAL1	44.8	3.7	0.701
37,796	7.36	57	IDH2	29.4	6.7	0.330
37,886	6.49	106	ILV5	76.0	4.5	0.892
38,700	7.83	55	BAT1	30.9	11.2	0.469
38,702	6.24	46	QCR2	NA[d]	2.2	0.326
39,477	5.58	86	FBA1	17.8	183.6	0.935
39,477	5.58	87	FBA1	427.2	183.6	0.935
39,540	6.50	150	HOM2	60.3	4.5	0.592
39,561	6.12	156	PSA1	96.4	27.5	0.718
41,158	6.01	49	YNL134C	14.9	1.5	0.316
41,623	7.18	58	BAT2	19.0	8.9	0.250
41,728	7.29	110	ERG10	24.1	4.5	0.543
41,900	5.42	74	TOM40	22.3	2.2	0.375
42,402	6.29	45	CYS3	6.7	8.9	0.621
42,883	5.63	67	DYS1	15.8	5.2	0.526
43,409	6.31	107	SER1	10.5	1.5	0.292
43,421	5.59	91	ERG6	2.2	14.1	0.408
44,174	7.32	56	YBR025C	13.1	6.0	0.684
44,682	4.99	72	TIF1	2.9	39.4	0.834
44,707	7.77	108	PGK1	23.7	165.7	0.897
44,707	7.77	109	PGK1	315.2	165.7	0.897
46,080	6.72	30	CAR2	15.4	NA[c]	0.495
46,383	8.52	53	IDP1	7.7	0.7	0.436
46,553	5.98	47	IDP2	32.4	NA[c]	0.197
46,679	6.39	50	ENO1	35.4	0.7	0.930
46,679	6.39	51	ENO1	6.6	0.7	0.930
46,679	6.39	52	ENO1	2.2	0.7	0.930
46,773	5.82	63	ENO2	15.5	289.1	0.960
46,773	5.82	64	ENO2	635.5	289.1	0.960
46,773	5.82	65	ENO2	93.0	289.1	0.960
46,773	5.82	66	ENO2	31.0	289.1	0.960
47,402	6.09	126	COR1	2.5	0.7	0.422
47,666	8.98	54	AAT2	11.7	6.0	0.338
48,364	5.25	73	WTM1	74.5	13.4	0.365
48,530	6.20	61	MET17	38.1	29.0	0.576
48,904	5.18	69	LYS9	16.2	3.7	0.463
48,987	4.90	153	SUP45	29.6	11.9	0.377
49,727	5.47	70	PRO2	13.6	5.2	0.297
49,912	9.27	62	TEF2	558.5	282.0	0.932
50,444	5.67	35	YDR190C	4.8	2.2	0.228
50,837	6.11	32	YEL047C	3.8	1.5	0.387
50,891	4.59	151	TUB2	11.2	7.4	0.404
51,547	6.80	27	LPD1	18.9	2.2	0.351
52,216	7.25	29	SHM2	19.7	7.4	0.722
52,859	5.54	37	YFR044C	30.2	6.7	0.442
53,798	5.19	71	HXK2	26.5	7.4	0.756
53,803	6.05	145	GYP6	4.4	0.7	0.147
54,403	5.29	39	ALD6	37.7	2.2	0.664
54,403	5.29	40	ALD6	6.6	2.2	0.664
54,502	6.20	31	ADE13	6.3	1.5	0.417
54,543	7.75	25	PYK1	225.3	101.8	0.965
54,543	7.75	26	PYK1	39.8	101.8	0.965
55,221	6.66	146	YEL071W	16.3	3.0	0.244
55,295	4.35	134	PDI1	66.2	14.1	0.589
55,364	5.98	24	GLK1	22.6	6.0	0.237
55,481	7.97	118	ATP1	21.6	2.2	0.637
55,886	6.47	28	CYS4	22.2	NA[c]	0.444
56,167	5.83	33	ARO8	14.3	3.0	0.324
56,167	5.83	34	ARO8	9.1	3.0	0.324
56,584	6.36	20	CYB2	18.9	NA[c]	0.259
57,366	5.53	60	FRS2	2.3	0.7	0.451
57,383	5.98	144	ZWF1	5.6	0.7	0.215
57,464	5.49	36	THR4	21.4	3.7	0.508
57,512	5.50	7	SRV2	6.5	NA[c]	0.260
57,727	4.92	152	VMA2	33.7	8.9	0.546
58,573	6.47	17	ACH1	4.4	1.5	0.327
58,573	6.47	18	ACH1	5.4	1.5	0.327
61,353	5.87	21	PDC1	6.5	200.7	0.962
61,353	5.87	22	PDC1	303.2	200.7	0.962
61,353	5.87	23	PDC1	16.3	200.7	0.962
61,649	5.54	38	CCT8	2.2	1.5	0.271

Continued

Continued on following page

TABLE 1—*Continued*

Mol wt	pI	Spot no.	YPD gene name[a]	Protein abundance (10^3 copies/cell)	mRNA abundance (copies/cell)	Codon bias
61,902	6.21	101	PDC5	4.3	NA[c]	0.828
62,266	6.19	16	ICL1	20.1	NA[c]	0.327
62,862	8.02	19	ILV3	5.3	4.5	0.548
63,082	6.40	119	PGM2	2.2	3.0	0.402
64,335	5.77	5	PAB1	30.4	1.5	0.616
66,120	5.42	8	STI1	6.7	0.7	0.313
66,120	5.42	9	STI1	6.4	0.7	0.313
66,450	5.29	141	SSB2	7.0	NA[c]	0.880
66,450	5.29	142	SSB2	2.3	NA[c]	0.880
66,456	5.23	10	SSB1	64.5	79.5	0.907
66,456	5.23	11	SSB1	59.0	79.5	0.907
66,456	5.23	12	SSB1	13.7	79.5	0.907
68,397	5.82	82	LEU4	3.1	3.0	0.407
69,313	4.90	13	SSA2	24.3	18.6	0.892
69,313	4.90	14	SSA2	77.1	18.6	0.892
74,378	8.46	15	YKL029C	2.8	3.7	0.353
75,396	5.82	6	GRS1	5.5	7.4	0.500
85,720	6.25	1	MET6	2.0	NA[c]	0.772
85,720	6.25	2	MET6	10.9	NA[c]	0.772
85,720	6.25	3	MET6	1.4	NA[c]	0.772
93,276	6.11	131	EFT1	17.9	41.6	0.890
93,276	6.11	132	EFT1	5.7	41.6	0.890
102,064[e]	6.61[e]	94	ADE3	4.8	5.2	0.423
107,482[e]	5.33[e]	4	MCM3	2.7	NA[c]	0.240

[a] YPD gene names are available from the YPD website (39).
[b] NA, calculation could not be performed or was not available.
[c] mRNA data inconclusive or NA.
[d] No methionines in predicted ORF; therefore, protein concentration was not determined.
[e] Measured molecular weight or pI did not match theoretical molecular weight or pI.

Protein quantitation. [^{35}S]methionine-labeled gels were exposed to X-ray film overnight, and then the silver stain and film were used to excise 156 spots of varying intensities, molecular weights, and pIs. The excised spots were placed in 0.6-ml microcentrifuge tubes, and scintillation cocktail (100 μl) was added. The samples were vortexed and counted. In addition, two parallel gels were electroblotted to polyvinylidene difluoride membranes. The membranes were exposed to X-ray film, and four intense single spots were excised from each membrane and subjected to amino acid analysis. For these four spots, a mean of 209 ± 4 cpm/pmol of protein/methionine was found. This number was used to quantitate all remaining spots in conjunction with the number of methionines present in the protein.

To ensure that proteins were labeled to equilibrium, parallel 2D gels were prepared and run on yeast metabolically labeled for 1, 2, 6, or 18 h. The corresponding 156 spots were excised from each gel, and radioactivity was measured by liquid scintillation counting for each spot. Calculated protein levels were highly reproducible for all time points measured after 1 h.

Calculation of codon bias and predicted half-life. Codon bias values were extracted from the YPD spreadsheet (17). Protein half-lives were calculated based on the N-end rule (33). When the N-terminal processing was not known experimentally, it was predicted based on the affinity of methionine aminopeptidase (31).

RESULTS

Characteristics of proteome approach. Nearly every facet of proteome analysis hinges on the unambiguous identification of large numbers of expressed proteins in cells. Several techniques have been described previously for the identification of proteins separated by 2DE, including N-terminal and internal sequencing (1, 2), amino acid analysis (38), and more recently mass spectrometry (25). We utilized techniques based on mass spectrometry because they afford the highest levels of sensitivity and provide unambiguous identification. The specific procedure used is schematically illustrated in Fig. 1 and is based on three principles. First, proteins are removed from the gel by proteolytic in-gel digestion, and the resulting peptides are separated by on-line capillary high-performance liquid chromatography. Second, the eluting peptides are ionized and detected, and the specific peptide ions are selected and fragmented by the mass spectrometer. To achieve this, the mass spectrometer switches between the MS mode (for peptide mass identification) and the MS/MS mode (for peptide characterization and sequencing). Selected peptides are fragmented by a process called collision-induced dissociation (CID) to generate a tandem mass spectrum (MS/MS spectrum) that contains the peptide sequence information. Third, individual CID mass spectra are then compared by computer algorithms to predicted spectra from a sequence database. This results in the identification of the peptide and, by association, the protein(s) in the spot. Unambiguous protein identification is attained in a single analysis by the detection of multiple peptides derived from the same protein.

Protein identification. Yeast total cell protein lysate (40 μg), metabolically labeled with [^{35}S]methionine, was electrophoretically separated by isoelectric focusing in the first dimension and by SDS–10% polyacrylamide gel electrophoresis in the second dimension. Proteins were visualized by silver staining and by autoradiography. Of the more than 1,000 proteins visible by silver staining, 156 spots were excised from the gel and subjected to in-gel tryptic digestion, and the resulting peptides were analyzed and identified by microspray LC-MS/MS techniques as described above. The proteins in this study were all identified automatically by computer software with no human interpretation of mass spectra. They are indicated in Fig. 2 and detailed in Table 1.

The CID spectra shown in Fig. 3 indicate that the quality of the identification data generated was suitable for unambiguous protein identification. The spectra represent the amino acid sequences of tryptic peptides NSGDIVNLGSIAGR (Fig. 3A) and FAVGAFTDSLR (Fig. 3B). Both peptides were derived from protein S57593 (hypothetical protein YMR226C), which migrated to spot 114 (molecular weight, 29,156; pI, 6.59) in the 2D gel in Fig. 2. Five other peptides from the same analysis were also computer matched to the same protein sequence.

Protein and mRNA quantitation. For the 156 genes investigated, the protein expression levels ranged from 2,200 (PGM2) to 863,000 (TDH2/TDH3) copies/cell. The levels of mRNA for each of the genes identified were calculated from SAGE frequency tables (35). These tables contain the mRNA levels for 4,665 genes in yeast strain YPH499 grown to mid-log phase in YPD medium on glucose as a carbon source. In some instances, the mRNA levels could not be calculated for reasons stated in Materials and Methods. For the proteins analyzed in this study, mean transcript levels varied from 0.7 to 473 copies/cell.

Selection of the sample population for mRNA-protein expression level correlation. The protein spots selected for identification were selected from spots visible by silver staining in the 2D gel. An attempt was made not to include spots where overlap with other spots was readily apparent. The number of proteins identified was 156 (Table 1). Some proteins migrated to more than one spot (presumably due to differential protein processing or modifications), and protein levels from these spots were calculated by integrating the intensities of the different spots. The 156 protein spots analyzed represented the products of 128 different genes. Genes were excluded from the correlation analysis only if part of the data set was missing; i.e., genes were excluded if (i) no mRNA expression data were available for the protein or putative SAGE tags were ambiguous, (ii) the amino acid sequence did not contain methionine, (iii) more than a single protein was conclusively identified as

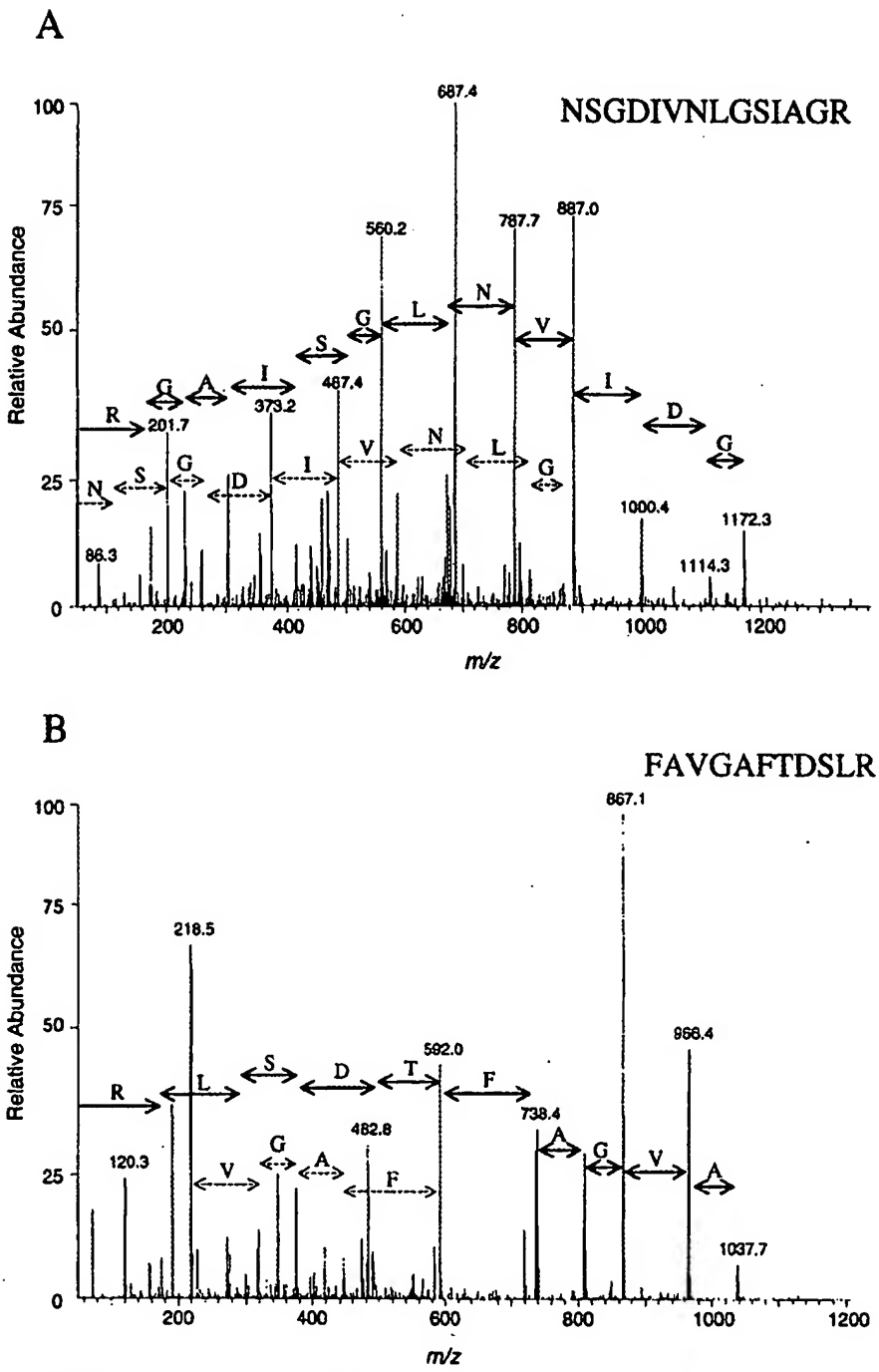


FIG. 3. Tandem mass (MS/MS) spectra resulting from analysis of a single spot on a 2D gel. The first quadrupole selected a single mass-to-charge ratio (*m/z*) of 687.2 (A) or 592.6 (B), while the collision cell was filled with argon gas, and a voltage which caused the peptide to undergo fragmentation by CID was applied. The third quadrupole scanned the mass range from 50 to 1,400 *m/z*. The computer program Sequest (8) was utilized to match MS/MS spectra to amino acid sequence by database searching. Both spectra matched peptides from the same protein, S57593 (yeast hypothetical protein YMR226C). Five other peptides from the same analysis were matched to the same protein.

migrating to the same gel spot, or (iv) the theoretical and observed pIs and molecular weights could not be reconciled. After these criteria were applied, the number of genes used in the correlation analysis was 106.

Codon bias and predicted half-lives. Codon bias is thought to be an indicator of protein expression, with highly expressed proteins having large codon bias values. The codon bias distribution for the entire set of more than 6,000 predicted yeast

gene ORFs is presented in Fig. 4A. The interval with the largest frequency of genes is between the codon bias values of 0.0 and 0.1. This segment contains more than 2,500 genes. The distribution of the codon bias values of the 128 different genes found in this study (all protein spots from Fig. 2) is shown in Fig. 4B, and protein half-lives (predicted from applying the N-end rule [33] to the experimentally determined or predicted protein N termini) are shown in Fig. 4C. No genes were identified with codon bias values less than 0.1 even though thousands of genes exist in this category. In addition, nearly all of the proteins identified had long predicted half-lives (greater than 30 h).

Correlation of mRNA and protein expression levels. The correlation between mRNA and protein levels of the genes selected as described above is shown in Fig. 5. For the entire group (106 genes) for which a complete data set was generated, there was a general trend of increased protein levels resulting from increased mRNA levels. The Pearson product moment correlation coefficient for the whole data set (106 genes) was 0.935. This number is highly biased by a small number of genes with very large protein and message levels. A more representative subset of the data is shown in the inset of Fig. 5. It shows genes for which the message level was below 10 copies/cell and includes 69% (73 of 106 genes) of the data used in the study. The Pearson product moment correlation coefficient for this data set was only 0.356. We also found that levels of protein expression coded for by mRNA with comparable abundance varied by as much as 30-fold and that the mRNA levels coding for proteins with comparable expression levels varied by as much as 20-fold.

The distortion of the correlation value induced by the uneven distribution of the data points along the x axis is further demonstrated by the analysis in Fig. 6. The 106 samples included in the study were ranked by protein abundance, and the Pearson product moment correlation coefficient was repeatedly calculated after including progressively more, and higher-abundance, proteins in each calculation. The correlation values remained relatively stable in the range of 0.1 to 0.4 if the lowest-expressed 40 to 95 proteins used in this study were included. However, the correlation value steadily climbed by the inclusion of each of the 11 very highly expressed proteins.

Correlation of protein and mRNA expression levels with codon bias. Codon bias is the propensity for a gene to utilize the same codon to encode an amino acid even though other codons would insert the identical amino acid in the growing polypeptide sequence. It is further thought that highly expressed proteins have large codon biases (3). To assess the value of codon bias for predicting mRNA and protein levels in exponentially growing yeast cells, we plotted the two experimental sets of data versus the codon bias (Fig. 7). The distribution patterns for both mRNA and protein levels with respect to codon bias were highly similar. There was high variability in the data within the codon bias range of 0.8 to 1.0. Although a large codon bias generally resulted in higher protein and message expression levels, codon bias did not appear to be predictive of either protein levels or mRNA levels in the cell.

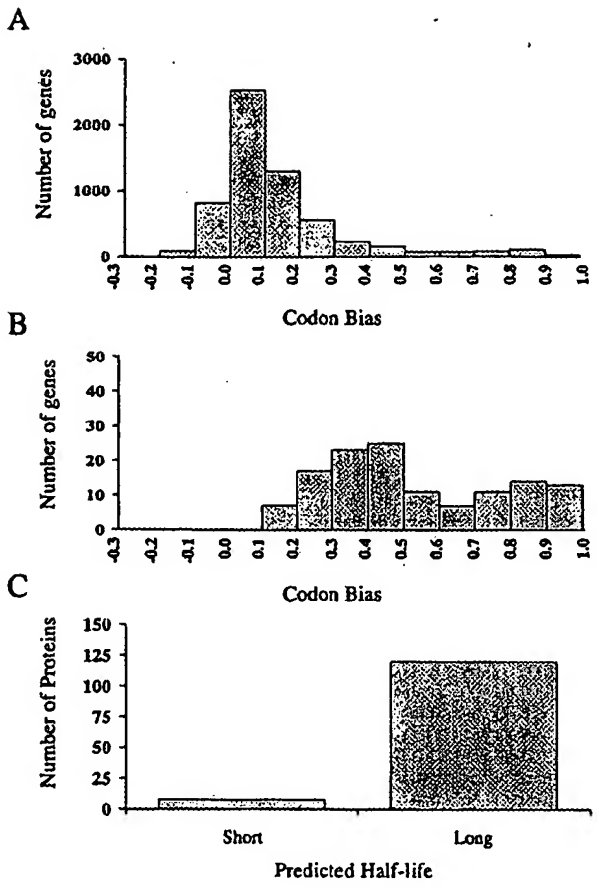


FIG. 4. Current proteome analysis technology utilizing 2DE without preenrichment samples mainly highly expressed and long-lived proteins. Genes encoding highly expressed proteins generally have large codon bias values. (A) Distribution of the yeast genome (more than 6,000 genes) based on codon bias. The interval with the largest frequency of genes is 0.0 to 0.1, with more than 2,500 genes. (B) Distribution of the genes from identified proteins in this study based on codon bias. No genes with codon bias values less than 0.1 were detected in this study. (C) Distribution of identified proteins in this study based on predicted half-life (estimated by N-end rule).

DISCUSSION

The desired end point for the description of a biological system is not the analysis of mRNA transcript levels alone but also the accurate measurement of protein expression levels and their respective activities. Quantitative analysis of global mRNA levels currently is a preferred method for the analysis of the state of cells and tissues (11). Several methods which either provide absolute mRNA abundance (34, 35) or relative mRNA levels in comparative analyses (20, 27) have been described elsewhere. The techniques are fast and exquisitely sensitive and can provide mRNA abundance for potentially any expressed gene. Measured mRNA levels are often implicitly or explicitly extrapolated to indicate the levels of activity of the corresponding protein in the cell. Quantitative analysis of protein expression levels (proteome analysis) is much more time-consuming because proteins are analyzed sequentially one by one and is not general because analyses are limited to the relatively highly expressed proteins. Proteome analysis does, however, provide types of data that are of critical importance for the description of the state of a biological system and that are not readily apparent from the sequence and the level of expression of the mRNA transcript. This study attempts to examine the relationship between mRNA and protein expression levels for a large number of expressed genes in cells representing the same state.

Limits in the sensitivity of current protein analysis technology precluded a completely random sampling of yeast proteins. We therefore based the study on those proteins visible by silver

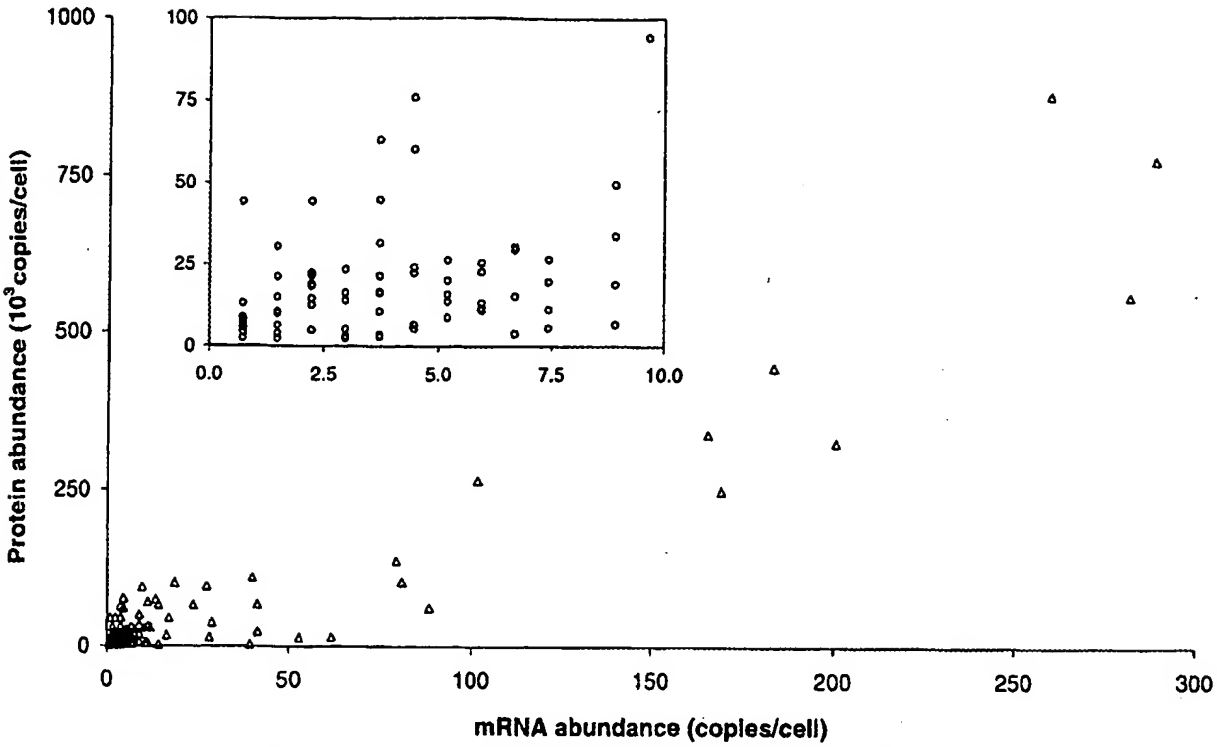


FIG. 5. Correlation between protein and mRNA levels for 106 genes in yeast growing at log phase with glucose as a carbon source. mRNA and protein levels were calculated as described in Materials and Methods. The data represent a population of genes with protein expression levels visible by silver staining on a 2D gel chosen to include the entire range of molecular weights, isoelectric focusing points, and staining intensities. The inset shows the low-end portion of the main figure. It contains 69% of the original data set. The Pearson product moment correlation for the entire data set was 0.935. The correlation for the inset containing 73 proteins (69%) was only 0.356.

staining on a 2D gel. Of the more than 1,000 visible spots, 156 were chosen to include the entire range of molecular weights, isoelectric focusing points, and staining intensities displayed on the 2D protein pattern. The genes identified in this study shared a number of properties. First, all of the proteins in this study had a codon bias of greater than 0.1 and 93% were greater than 0.2 (Fig. 4B). Second, with few exceptions, the proteins in this study had long predicted half-lives according to the N-end rule (Fig. 4C). Third, low-abundance proteins with regulatory functions such as transcription factors or protein kinases were not identified.

Because the population of proteins used in this study appears to be fairly homogeneous with respect to predicted half-life and codon bias, it might be expected that the correlation of the mRNA and protein expression levels would be stronger for this population than for a random sample of yeast proteins. We tested this assumption by evaluating the correlation value if different subsets of the available data were included in the calculation. The 106 proteins were ranked from lowest to highest protein expression level, and the trend in the correlation value was evaluated by progressively including more of the higher-abundance proteins in the calculation (Fig. 6). The correlation value when only the lower-abundance 40 to 93 proteins were examined was consistently between 0.1 and 0.4. If the 11 most abundant proteins were included, the correlation steadily increased to 0.94. We therefore expect that the correlation for all yeast proteins or for a random selection would be less than 0.4. The observed level of correlation between mRNA and protein expression levels suggests the importance of posttranslational mechanisms controlling gene expression. Such mechanisms include translational control (15) and control of protein half-life (33). Since these mechanisms are also active in higher eukaryotic cells, we speculate that there is no predictive correlation between steady-state levels of mRNA and those of protein in mammalian cells.

Like other large-scale analyses, the present study has several potential sources of error related to the methods used to determine mRNA and protein expression levels. The mRNA levels were calculated from frequency tables of SAGE data. This method is highly quantitative because it is based on actual sequencing of unique tags from each gene, and the number of times that a tag is represented is proportional to the number of mRNA molecules for a specific gene. This method has some limitations including the following: (i) the magnitude of the error in the measurement of mRNA levels is inversely proportional to the mRNA levels, (ii) SAGE tags from highly similar genes may not be distinguished and therefore are summed, (iii) some SAGE tags are from sequences in the 3′ untranslated region of the transcript, (iv) incomplete cleavage at the SAGE tag site by the restriction enzyme can result in two tags representing one mRNA, and (v) some transcripts actually do not generate a SAGE tag (34, 35).

For the SAGE method, the error associated with a value increases with a decreasing number of transcripts per cell. The conclusions drawn from this study are dependent on the quality of the mRNA levels from previously published data (35). Since more than 65% of the mRNA levels included in this study were calculated to 10 copies/cell or less (40% were less

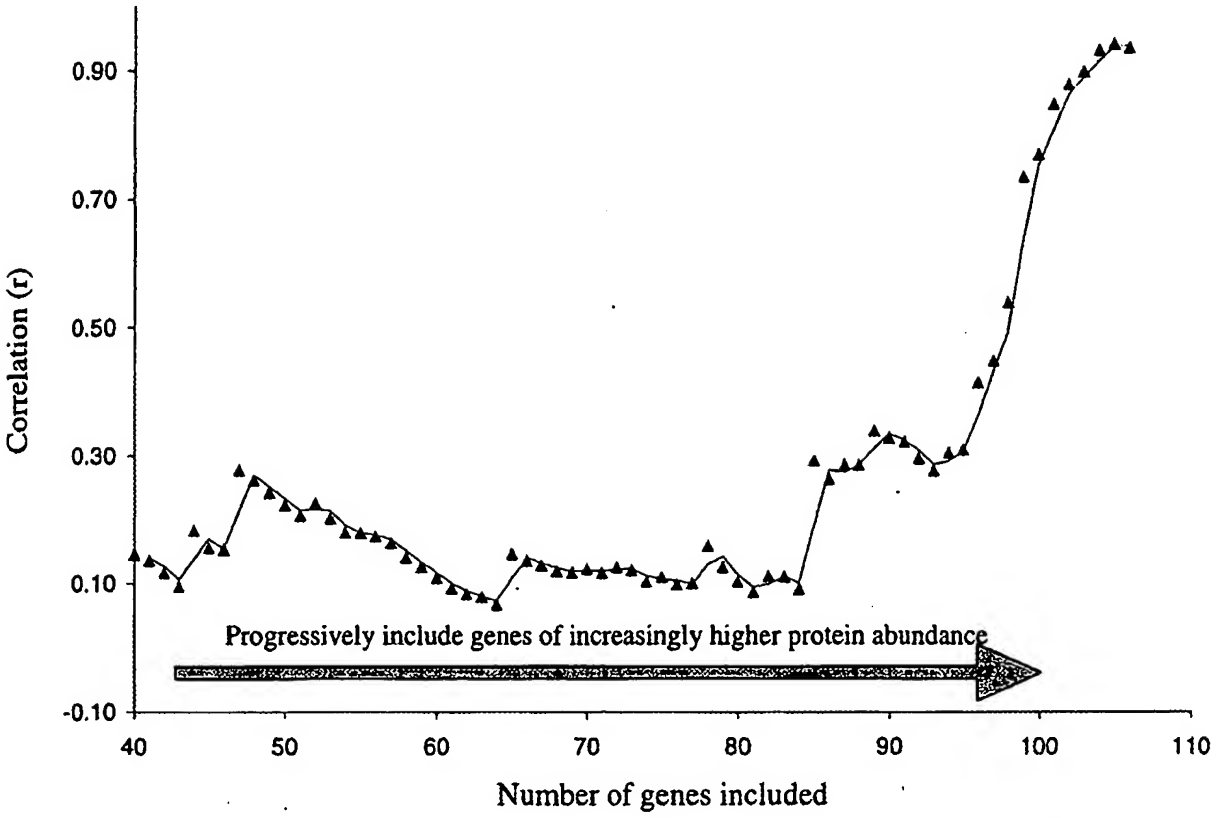


FIG. 6. Effect of highly abundant proteins on Pearson product moment correlation coefficient for mRNA and protein abundance in yeast. The set of 106 genes was ranked according to protein abundance, and the correlation value was calculated by including the 40 lowest-abundance genes and then progressively including the remaining 66 genes in order of abundance. The correlation value climbs as the final 11 highly abundant proteins are included.

than 4 copies/cell), the error associated with these values may be quite large. The mRNA levels were calculated from more than 20,000 transcripts. Assuming that the estimate of 15,000 mRNA molecules per cell is correct (16), this would mean that mRNA transcripts present at only a single copy per cell would be detected 72% of the time (35). The mRNA levels for each gene were carefully scrutinized, and only mRNA levels for which a high degree of confidence existed were included in the correlation value.

Protein abundance was determined by metabolic radiolabeling with [^{35}S]methionine. The calculation required knowledge of three variables: the number of methionines in the mature protein, the radioactivity contained in the protein, and the specific activity of the radiolabel normalized per methionine. The number of methionines per protein was determined from the amino acid sequence of the proteins identified by tandem mass spectrometry. For some proteins, it was not known whether the methionine of the nascent polypeptide was processed away. The N termini of those proteins were predicted based on the specificity of methionine aminopeptidase (31). If the N-terminal processing did not conform to the predicted specificity of processing enzymes, the calculation of the number of methionines would be affected. This discrepancy would affect most the quantitation of a protein with a very low number of methionines. The average number of calculated methionines per protein in this study was 7.2. We therefore expect the potential for erroneous protein quantitation due to unusual N-terminal processing to be small.

The amount of radioactivity contained in a single spot might be the sum of the radioactivity of comigrating proteins. Because protein identification was based on tandem mass spectrometric techniques, comigrating proteins could be identified. However, comigrating proteins were rarely detected in this study, most likely because relatively small amounts of total protein (40 μg) were initially loaded onto the gels, which resulted in highly focused spots containing generally 1 to 25 ng of protein. Because of the relatively small amount loaded, the concentrations of any potentially comigrating protein would likely be below the limit of detection of the mass spectrometry technique used in this study (1 to 5 ng) and below the limit of visualization by silver staining (1 to 5 ng). In the overwhelming majority of the samples analyzed, numerous peptides from a single protein were detected. It is assumed that any comigrating proteins were at levels too low to be detected and that their influence in the calculation would be small.

The specific activity of the radiolabel was determined by relating the precise amount of protein present in selected spots of a parallel gel, as determined by quantitative amino acid composition analysis, to the number of methionines present in the sequence of those proteins and the radioactivity determined by liquid scintillation counting. It is possible that the resulting number might be influenced by unavoidable losses inherent in the amino acid analysis procedure applied. Because four different proteins were utilized in the calculation and the experiment was done in duplicate, the specific activity calculated is thought to be highly accurate. Indeed, the specific

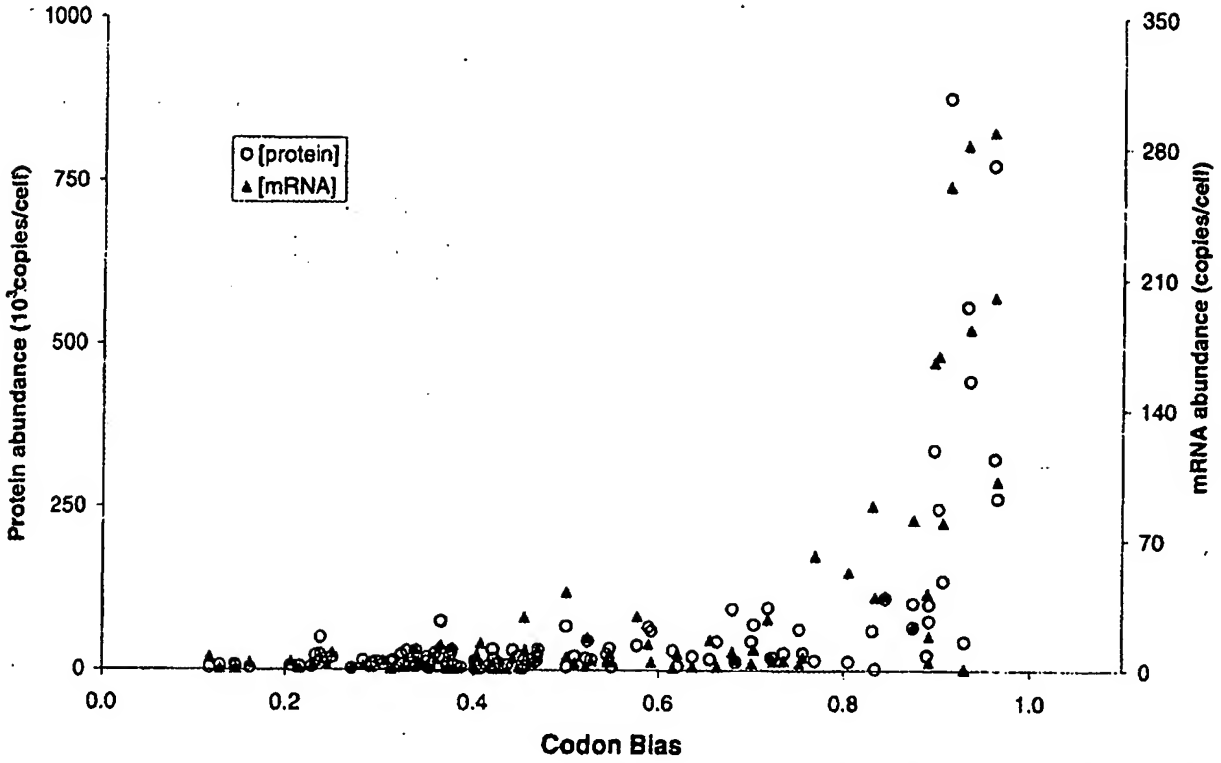


FIG. 7. Relationship between codon bias and protein and mRNA levels in this study. Yeast mRNA and protein expression levels were calculated as described in Materials and Methods. The data represent the same 106 genes as in Fig. 5.

activities calculated for each of the four proteins varied by less than 10%. Any inconsistencies in the calculation of the specific activity would result in differences in the absolute levels calculated but not in the relative numbers and would therefore not influence the correlation value determined.

The protein quantitative method used eliminates a number of potential errors inherent in previous methods for the quantitation of proteins separated by 2DE, such as preferential protein staining and bias caused by inequalities in the number of radiolabeled residues per protein. Any 2D gel-based method of quantitation is complicated by the fact that in some cases the translation products of the same mRNA migrated to different spots. One major reason is posttranslational modification or processing of the protein. Also, artifactual proteolysis during cell lysis and sample preparation can lead to multiple resolved forms of the protein. In such cases, the protein levels of spots coded for by the same mRNA were pooled. In addition, the existence of other spots coded for by the same mRNA that were not analyzed by mass spectrometry or that were below the limit of detection for silver staining cannot be ruled out. However, since this study is based on a class of highly expressed proteins, the presence of undetected minor spots below silver staining sensitivity corresponding to a protein analyzed in the study would generally cause a relatively small error in protein quantitation.

Codon bias is a measure of the propensity of an organism to selectively utilize certain codons which result in the incorporation of the same amino acid residue in a growing polypeptide chain. There are 61 possible codons that code for 20 amino acids. The larger the codon bias value, the smaller the number of codons that are used to encode the protein (19). It is thought that codon bias is a measure of protein abundance because highly expressed proteins generally have large codon bias values (3, 13).

Nearly all of the most highly expressed proteins had codon bias values of greater than 0.8. However, we detected a number of genes with high codon bias and relative low protein abundance (Fig. 7). For example, the expressed gene with both the second largest protein and mRNA levels in the study was ENO2_YEAST (775,000 and 289.1 copies/cell, respectively). ENO1_YEAST was also present in the gel at much lower protein and mRNA levels (44,200 and 0.7 copies/cell, respectively). The codon bias values for ENO2 and ENO1 are similar (0.96 and 0.93, respectively), but the expression of the two genes is differentially regulated. Specifically, ENO1_YEAST is glucose repressed (6) and was therefore present in low abundance under the conditions used. Other genes with large codon bias values that were not of high protein abundance in the gel include EFT1, TIF1, HXK2, GSP1, EGD2, SHM2, and TAL1. We conclude that merely determining the codon bias of a gene is not sufficient to predict its protein expression level.

Interestingly, codon bias appears to be an excellent indicator of the boundaries of current 2D gel proteome analysis technology. There are thousands of genes with expressed mRNA and likely expressed protein with codon bias values less than 0.1 (Fig. 4A). In this study, we detected none of them, and only a very small percentage of the genes detected in this study had codon bias values between 0.1 and 0.2 (Fig. 4B). Indeed, in every examined yeast proteome study (5, 7, 13, 28) where the combined total number of identified proteins is 300 to 400, this same observation is true. It is expected that for the more complex cells of higher eukaryotic organisms the detection of

low-abundance proteins would be even more challenging than for yeast. This indicates that highly abundant, long-lived proteins are overwhelmingly detected in proteome studies. If proteome analysis is to provide truly meaningful information about cellular processes, it must be able to penetrate to the level of regulatory proteins, including transcription factors and protein kinases. A promising approach is the use of narrow-range focusing gels with immobilized pH gradients (IPG) (23). This would allow for the loading of significantly more protein per pH unit covered and also provide increased resolution of proteins with similar electrophoretic mobilities. A standard pH gradient in an isoelectric focusing gel covers a 7-pH-unit range (pH 3 to 10) over 18 cm. A narrow-range focusing gel might expand the range to 0.5 pH units over 18 cm or more. This could potentially increase by more than 10-fold the number of proteins that can be detected. Clearly, current proteome technology is incapable of analyzing low-abundance regulatory proteins without employing an enrichment method for relatively low-abundance proteins. In conclusion, this study examined the relationship between yeast protein and message levels and revealed that transcript levels provide little predictive value with respect to the extent of protein expression.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Science and Technology Center for Molecular Biotechnology, NIH grant T32HG00035-3, and a grant from Oxford Glycosciences.

We thank Jimmy Eng for expert computer programming, Garry Corthals and John R. Yates III for critical discussion, and Siavash Mohandesi for expert technical help.

REFERENCES

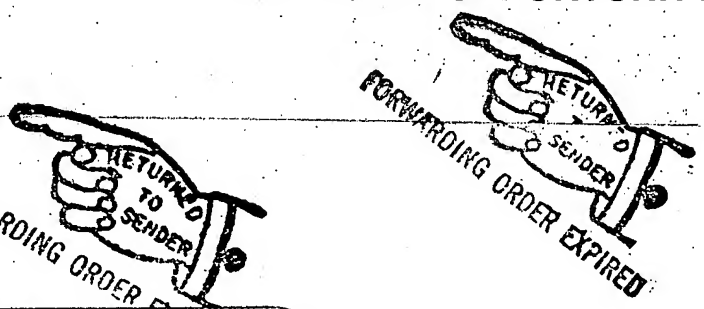
1. Aebersold, R. H., J. Leavitt, R. A. Saavedra, L. E. Hood, and S. B. Kent. 1987. Internal amino acid sequence analysis of proteins separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose. Proc. Natl. Acad. Sci. USA 84:6970–6974.
2. Aebersold, R. H., D. B. Teplow, L. E. Hood, and S. B. Kent. 1986. Electroblotting onto activated glass. High efficiency preparation of proteins from analytical sodium dodecyl sulfate-polyacrylamide gels for direct sequence analysis. Eur. J. Biochem. 161:4229–4238.
3. Bennetzen, J. L., and B. D. Hall. 1982. Codon selection in yeast. J. Biol. Chem. 257:3026–3031.
4. Boucherie, H., G. Dujardin, M. Kermorgant, C. Monribot, P. Slonimski, and M. Perrot. 1995. Two-dimensional protein map of Saccharomyces cerevisiae: construction of a gene-protein index. Yeast 11:601–613.
5. Boucherie, H., F. Sagliocco, R. Joubert, I. Maillet, J. Labarre, and M. Perrot. 1996. Two-dimensional gel protein database of Saccharomyces cerevisiae. Electrophoresis 17:1683–1699.
6. Carmen, A. A., P. K. Brindle, C. S. Park, and M. J. Holland. 1995. Transcriptional regulation by an upstream repression sequence from the yeast enolase gene ENO1. Yeast 11:1031–1043.
7. Ducret, A., I. VanOostveen, J. K. Eng, J. R. Yates, and R. Aebersold. 1998. High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry. Protein Sci. 7:706–719.
8. Eng, J., A. McCormack, and J. R. Yates. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5:976–989.
9. Figeys, D., A. Ducret, J. R. Yates, and R. Aebersold. 1996. Protein identification by solid phase microextraction-capillary zone electrophoresis-microelectrospray-tandem mass spectrometry. Nat. Biotechnol. 14:1579–1583.
10. Figeys, D., I. VanOostveen, A. Ducret, and R. Aebersold. 1996. Protein identification by capillary zone electrophoresis/microelectrospray ionization-tandem mass spectrometry at the subfemtomole level. Anal. Chem. 68:1822–1828.
11. Fraser, C. M., and R. D. Fleischmann. 1997. Strategies for whole microbial genome sequencing and analysis. Electrophoresis 18:1207–1216.
12. Garrels, J. I., B. Futcher, R. Kobayashi, G. I. Latter, B. Schwender, T. Volpe, J. R. Warner, and C. S. McLaughlin. 1994. Protein identifications for a Saccharomyces cerevisiae protein database. Electrophoresis 15:1466–1486.
13. Garrels, J. I., C. S. McLaughlin, J. R. Warner, B. Futcher, G. I. Latter, R. Kobayashi, B. Schwender, T. Volpe, D. S. Anderson, F. Mesquita-Fuentes, and W. E. Payne. 1997. Proteome studies of Saccharomyces cerevisiae: identification and characterization of abundant proteins. Electrophoresis 18:1347–1360.
14. Gygi, S. P., and R. Aebersold. 1998. Absolute quantitation of 2-DE protein spots, p. 417–421. *In* A. J. Link (ed.), 2-D protocols for proteome analysis. Humana Press, Totowa, N.J.
15. Harford, J. B., and D. R. Morris. 1997. Post-transcriptional gene regulation. Wiley-Liss, Inc., New York, N.Y.
16. Hereford, L. M., and M. Rosbash. 1977. Number and distribution of polyadenylated RNA sequences in yeast. Cell 10:453–462.
17. Hodges, P. E., W. E. Payne, and J. I. Garrels. 1998. The Yeast Protein Database (YPD): a curated proteome database for Saccharomyces cerevisiae. Nucleic Acids Res. 26:68–72.
18. Klose, J., and U. Kobalz. 1995. Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome. Electrophoresis 16:1034–1059.
19. Kurland, C. G. 1991. Codon bias and gene expression. FEBS Lett. 285:165–169.
20. Lashkari, D. A., J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis. 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc. Natl. Acad. Sci. USA 94:13057–13062.
21. Liang, P., and A. B. Pardee. 1992. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257:967–971.
22. Link, A. J., L. G. Hays, E. B. Carmack, and J. R. Yates III. 1997. Identifying the major proteome components of Haemophilus influenzae type-strain NCTC 8143. Electrophoresis 18:1314–1334.
23. Nawrocki, A., M. R. Larsen, A. V. Podtelejnikov, O. N. Jensen, M. Mann, P. Roepstorff, A. Gorg, S. J. Fey, and P. M. Larsen. 1998. Correlation of acidic and basic carrier ampholyte and immobilized pH gradient two-dimensional gel electrophoresis patterns based on mass spectrometric protein identification. Electrophoresis 19:1024–1035.
24. O'Farrell, P. H. 1975. High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. 250:4007–4021.
24a. OWL Protein Sequence Database. 2 August 1998, posting date. [Online.] http://bmbsgi11.leeds.ac.uk/bmb5dp/owl.html. [8 January 1999, last date accessed.]
25. Patterson, S. D., and R. Aebersold. 1995. Mass spectrometric approaches for the identification of gel-separated proteins. Electrophoresis 16:1791–1814.
26. Pennington, S. R., M. R. Wilkins, D. F. Hochstrasser, and M. J. Dunn. 1997. Proteome analysis: from protein characterization to biological function. Trends Cell Biol. 7:168–173.
27. Shalon, D., S. J. Smith, and P. O. Brown. 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. 6:639–645.
28. Shevchenko, A., O. N. Jensen, A. V. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, H. Boucherie, and M. Mann. 1996. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. Proc. Natl. Acad. Sci. USA 93:14440–14445.
29. Shevchenko, A., M. Wilm, O. Vorm, and M. Mann. 1996. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. Anal. Chem. 68:850–858.
30. Sikorski, R. S., and P. Hieter. 1989. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. Genetics 122:19–27.
31. Tsunasawa, S., J. W. Stewart, and F. Sherman. 1985. Amino-terminal processing of mutant forms of yeast iso-1-cytochrome c. The specificities of methionine aminopeptidase and acetyltransferase. J. Biol. Chem. 260:5382–5391.
32. Urlinger, S., K. Kuchler, T. H. Meyer, S. Uebel, and R. Tamp'e. 1997. Intracellular location, complex formation, and function of the transporter associated with antigen processing in yeast. Eur. J. Biochem. 245:266–272.
33. Varshavsky, A. 1996. The N-end rule: functions, mysteries, uses. Proc. Natl. Acad. Sci. USA 93:12142–12149.
34. Velculescu, V. E., L. Zhang, B. Vogelstein, and K. W. Kinzler. 1995. Serial analysis of gene expression. Science 270:484–487.
35. Velculescu, V. E., L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett, Jr., P. Hieter, B. Vogelstein, and K. W. Kinzler. 1997. Characterization of the yeast transcriptome. Cell 88:243–251.
36. Wilkins, M. R., K. L. Williams, R. D. Appel, and D. F. Hochstrasser. 1997. Proteome research: new frontiers in functional genomics. Springer-Verlag, Berlin, Germany.
37. Wilm, M., A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer, T. Fotsis, and M. Mann. 1996. Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. Nature 379:466–469.
38. Yan, J. X., M. R. Wilkins, K. Ou, A. A. Gooley, K. L. Williams, J. C. Sanchez, O. Golaz, C. Pasquali, and D. F. Hochstrasser. 1996. Large-scale amino-acid analysis for proteome studies. J. Chromatogr. A 736:291–302.
39. YPD Website. 6 March 1998, revision date. [Online.] Proteome, Inc. http://www.proteome.com/YPDhome.html. [8 January 1999, last date accessed.]

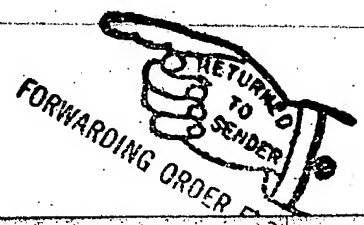
Organization TC1600 Bldg./Room REMSEN

U. S. DEPARTMENT OF COMMERCE
COMMISSIONER FOR PATENTS
P.O. BOX 1450
ALEXANDRIA, VA 22313-1450
IF UNDELIVERABLE RETURN IN TEN DAYS

OFFICIAL BUSINESS

AN EQUAL OPPORTUNIT





This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.